Consented High-performance Indexing and Retrieval of Pathology Specimens (CHIRPS)

July 2000

Harvard University
Boston, Massachusetts

CHIRPS

DESCRIPTION:

Pathology specimens and their associated clinical data, archived in repositories of various configurations, represent a vast and underutilized mine of valuable resource. The emergence of the Internet and its related technology now-offers an opportunity to coordinate these valuable resources. Cost, logistical, and public policy issues make a centralized repository of specimens and/or specimen-related information unpalatable. In "Consented High-performance Indexing and Retrieval of Pathology Specimens" (CHIRPS) we propose an approach to a distributed specimen-informatics network, that will allow institutions to maintain local control of specimens, related identifiers and other sensitive information while safely sharing anonymized data across institutions. CHIRPS will support a novel peer-to-peer network where each site will announce their presence to others, and will distribute queries among themselves, in a manner similar to GnutellaNet. One key addition to the model is a method for secure authentication of clients and servers when needed, through the use of digital certificates. These major goals will be addressed: 1) Establishing a scalable Extensible Markup Language (XML) representation of specimen annotation that will support both a least common denominator access and an advanced query access to existing specimen information across multiple healthcare delivery and research institutions, 2) Formulating a taxonomy of patient consent that is part of the XML-based annotation, allowing for a balance between the advancement of biomedical knowledge and the protection of patient privacy, and 3) Developing a peer-to-peer distributed architecture for indexing and searching for specimens that leverages the Internet and the Web, and that minimizes the effort needed to participate in the Shared Pathology Informatics Network (SPIN).

CHIRPS will be implemented and tested for its scalability across the Harvard/UCLA consortium, which joins two large academic medical centers, each with an established comprehensive cancer center, the Dana-Farber/Harvard Cancer Center and the Jonsson Comprehensive Cancer Center. The Harvard/UCLA consortium is composed of 9 member institutions representing 7 different pathology information systems and their associated archives containing millions of annotated specimens. The development and implementation of CHIRPS will provide a means to harness these existing valuable resources and a generalized platform to index and host specimen-related information prospectively to support future collaborations.

PERFORMANCE SITES:

Harvard Medical School Affiliates, Boston, MA:
Beth Israel Deaconess Medical Center (BID), Brigham & Women's Hospital (BWH), Children's Hospital (CH), Dana-Farber Cancer Institute (DFCI), MassachusettsGeneral Hospital (MGH)

UCLA and Affiliates, Los Angeles, CA:
UCLA Medical Center (UCLA), Olive View Medical Center (OVMC),
Cedars-Sinai Medical Center (CSMC), VA Greater LA Healthcare System (VAGLAS).

KEY PERSONNEL:

Name	Organization	Role on Project
Kohane, Isaac	CH	PI
Fletcher, Christopher	BWH	Co-PI, Chair, Steering Committee
Weeks, Jane	DFCI	Co-PI, Outcomes Research

Braun, Jonathan	UCLA	Co-PI, Chair, UCLA Steering Subcommittee
Chueh, Henry	MGH	Co-PI, Leader, Informatics Task Force
Beckwith, Bruce	BID	Co-PI, Member, Informatics Task Force
Greenes, Robert	BWH	Chair, Pathology-Inform. Advisory Board
Anthony, Douglas	CH	Member, Steering Committee
Halamka, John	BID	Co-PI, Member, Informatics Task Force
Drake, Thomas	UCLA	Member, Pathology-Inform. Advisory Board
McCoy, Michael	UCLA	Member, Steering Committee
Murphy, Shawn	MGH	Member, Informatics Task Force
Balis, Ulysses	MGH	Member, Informatics Task Force
Goodspeed, Barrett	MGH	Member, Informatics Task Force
Boxwala, Aziz	BWH	Member, Informatics Task Force
Kuperman, Gilad	BWH	Member, Informatics Task Force
Zeng, Qing	BWH	Member, Informatics Task Force
Roberts, Alice	UCLA	Member, Informatics Task Force

TABLE OF CONTENTS

A. Specific Aims

The title of the application is "Consented High-Performance Indexing and Retrieval of Pathology Specimens" (CHIRPS). CHIRPS will be designed and implemented within the context of the Departments of Pathology in each of the major Harvard Medical School-affiliated hospitals as well as those in the affiliated-hospitals of the University of California in Los Angeles. These highly heterogeneous systems (both organizationally and technologically) will serve as an in depth, scaleable testbed for deployment to larger numbers of institutions. The tools, methodologies and policies developed for CHIRPS will be made available on the Internet to all interested parties. In particular all software will be released within the highly successful framework of the Open Source (see www.opensource.org) agreement to ensure maximal dissemination with the least commercial barriers. The proposal will focus on three specific aims:

A.1. Define a scalable and extensible representation for tissue specimen annotation in Extensible Markup Language (XML) that can be queried effectively using Extensible Query Language (XQL).

A consistent method of annotating existing specimens will form the foundation of an informatics network to help share these specimens cooperatively. XML will be used as an Internet-enabled syntax for the annotation. Within the proposed consortium, there is minimal field-level information that is common between specimen databases. Consequently the approach to defining a specimen annotation must be extensible because the initial common elements are minimal.

A.2. Formulate a taxonomy for confidentiality and patient consent that describes how a specimen can be used**.**

Protecting the rights of patients must be balanced with providing safe and reasonable access to human material. Since the understanding of this balance remains a discussion, there is no one correct way to handle this issue at the moment. We propose the creation of a taxonomy for expressing varying degrees of confidentiality and consent that can be used to identify the potential use of individual specimens. This taxonomy will be implemented explicitly as part of the specimen annotation definition. It can be refined as the public policy discussions evolve.

A.3. Design a distributed architecture for indexing and searching for specimens that supports investigator-directed queries with a standard Web browser on the Internet.

A software architecture of peer-to-peer communication will form the basis of a scalable system of query transmission and results-set formation. This architecture will reflect the highly distributed model of the Gnutella Network.[1] Although trusted site authentication and management of standard metadata collections for query construction will occur at specific sites, the most substantive task of diffusing the queries throughout the system and collecting the result-set will be accomplished through a web of peers. Each CHIRPS server will function as a node capable of cooperating in this web.

B. Background and Significance

Modern biomedical research requires access to patient materials that includes tissues, diagnostic specimens and their related clinical data. Investigators within the hospital community have recognized the value of these specimens. Tissues in particular have been collected for many years and used effectively to advance various areas of biomedical research. For example, tissue banks have been historically valuable in the study of human cancer. With the imminent completion of the human genome project and the revolution in molecular genetic technology, these existing tissues become increasingly important as valuable sources of information. New molecular technologies are also increasingly applicable to formalin-fixed paraffin-embedded tissues, once only the domain of scarce fresh frozen tissues. This translates into potentially large-scale research on the millions of conventionally processed tissue specimens archived in anatomic pathology departments. However, practical access to large, diverse collections of annotated pathology specimens and related clinical data across institutional boundaries is very limited, at best. There are isolated specimen information databases that have been created within the context of retrospective or prospective studies of specific diseases such as prostate carcinoma or melanoma.[2,3] Based on these needs, the National Cancer Institute (NCI) has sponsored the

initiation of a number of tissue resources focused on certain tissue types and disease processes, such as the Cooperative Human Tissue Network, Breast Cancer Tissue Resources, the Clinical Trials Cooperative Groups and the AIDS and Cancer Specimen Bank.[4] Valuable as these resources are to the specific research areas, they are limited in scope and their maintenance in the current configuration is not cost-effective. To find an alternative approach, the NCI announced the current SPIN initiative targeting archived specimen and related clinical treatment and outcome data stored in the hospital pathology departments.

B.1. Informed by the proliferation of biomedical and bioinformatics databases

In contrast to the paucity of specimen databases, numerous biomedical databases have been created as central database repositories to which different investigators can contribute. The Collaborative Computational Project's bioinformatics resource Web page has a partial listing of 183 such databases.[17] The approach to the construction of these databases is often a model that consists of core datum annotated by descriptive information. For example, in the Tumor Gene Index the genes form the primary data, and they are annotated with associated information such as the tissue, stage of cancer and their pattern of expression in tissues. As part of the Cancer Genome Anatomy Project initiative sponsored by the NCI, the goal is to generate a catalog of annotated genes.[5,6] These annotations serve to provide a richness to the data that makes the databases suitable for searching by different characteristics.

However, most bioinformatics/biomedical databases contain minimal clinical information, and have lost their link to any individuals, if there ever were such links. Consequently the simplest approach of starting a central database for each topic area has proliferated. Since each database defines its own schemata and representation of the data being stored, this approach avoids the need to come to consensus agreement on how to represent even similar datatypes. Those familiar with these myriad individual databases recognize that even for this type of information more distributed methods will be needed.[7,8,9] Berman points out succinctly that the "…goal here would be to be able to query across large numbers of databases in a seamless way. This would replace the concept of putting everything into one large biological database or in using static cross-links…[so] that each database creator and each view of the information could be fully exploited and explored. This will necessitate further development of methods to organize and query such diverse information across networks. This requires new research and development in the fundamental computer infrastructure, a point that is often lost in discussions…".[10] Managing the complexities of having data that remains closely linked to individuals and hospitals makes a scalable approach to pathology specimen databases absolutely critical from the beginning.

B.2. Tapping existing data trapped in narrative reports

Unlike the often-terse language of genotypes, much of the information related to pathology specimens remains trapped in narrative reports. Natural language parsing (NLP) has been an effective approach to extracting noun phrases and basic concepts from narrative reports in narrow domains such as pathology.[11] It has been less successful in identifying higher level concepts and clinical context. Researchers have experimented with auto coding of pathology reports with natural language processing (NLP) and vocabulary mapping techniques. Coles, Dunham, Myers, Lamson et. al. developed programs to auto code pathology reports in 1970s. In 1990s, new pathology NLP applications were developed and evaluated. For example, Moore reported in 1994 that their NLP application for pathology performed better than manual coding.[12] Other more general medical NLP applications have also been used to code pathology reports. For instance, Friedman and et al developed and evaluated a NLP application for Mammography and chest X-ray reports.[13] This application was later extended to other medical domains including pathology. NLP typically consists of preprocessing, morphological and lexical processing, syntactic processing, and semantic analysis. During the past decade, information extraction (IE) has emerged as a new NLP area that focuses on the information processing needs associated with large volumes of text documents. Instead of full analysis of the meaning of a text, it extracts predefined, specific, structured information from text documents. The nature of these techniques makes them well suited for application against the volume of existing pathology reports.

B.3. Public policy issues of centralized clinical databases

Patients, investigators, hospital and institutional administrations, and outside regulatory agencies all expect appropriate use and distribution of tissue specimens.[14] Part of the administrative simplification section of HIPAA (Health Insurance Portability and Accountability Act) of 1996 is specifically concerned with inter-institutional communications, patient consent, patient identification and standards for ensuring compliance and protection of confidentiality without disruption of healthcare processes. The concerns articulated during the passage of HIPAA and the subsequent controversies surrounding its implementation, particularly as regards to patient consent and identification must inform the design of a SPIN architecture. It has clearly influenced the highly distributed and secure CHIRPS design. The investigators of CHIRPS have extensive experience in these HIPAA-related issues and have been working on National Library of Medicine-funded projects specifically addressing the challenges of identification and anonymization. There are complex ethical and academic issues associated with the acquisition, ownership and utilization of these specimens[15] and some of these issues are addressed within the CHIRPS proposal.

B.4. The significance of creating an infrastructure to share

Formalin-fixed paraffin-embedded tissue-based research has been undertaken principally in departments of pathology. Histologic and immunohistochemical techniques in these laboratories form the basis of how new tumor types are recognized, how classifications are refined, and how prognostic or diagnostic parameters are defined. This work represents the cornerstone of clinical research in academic surgical pathology, and it is absolutely dependent on the ability to retrieve properly coded specimens.[16] However, the rapid adoption of the aforementioned molecular techniques has greatly broadened the parties interested in appropriately coded specimens. Often researchers need a suitably large number of specimens of a given disease entity to produce biostatistically meaningful results in a study. In single institutions, specimens from common diseases such as carcinomas of lung, breast or prostate may be prevalent -- but even then it may be difficult to accumulate enough cases of a homogeneous type (e.g., of similar patient age, histologic grade, tumor stage). The problem is compounded with rare diseases and tumors, or in cases where only small amounts of materials are collected (e.g., brain biopsies). Consequently any practical initiative that can increase the size and accessibility of the specimen pool can have a major impact to the research community worldwide.[26] The successful implementation of SPIN will increase the value of existing specimen banks exponentially, enhancing the biomedical research leading to advances in patient care and quality of life.

C. Preliminary Studies

This section is divided into 1) the rationale for composition of the consortium, 2) a survey of the current information systems used by the multiple pathology laboratories in the CHIRPS consortium, and 3) relevant prior work relating to the key informatics technologies required of CHIRPS, particularly distributed querying over the web, cryptographic identification systems and analytic data warehouses.

C.1. Institutions of the Harvard/UCLA consortium & rationale for the composition

The Harvard/UCLA consortium is composed of 9 participating hospitals and research institutions, these being the affiliated institutions of 2 major academic medical centers, each having a comprehensive cancer center, the Dana Farber/Harvard Cancer Center (DF/HCC) and Jonsson Comprehensive Cancer Center, (JCCC), respectively, through which cancer research is coordinated. The 5 Harvard affiliates are members of the DF/HCC, a consortium cancer center initiated in October 1997. The Harvard affiliates include the Beth Israel Deaconess Hospital (BID), Brigham and Women's Hospital (BWH), Children's Hospital (CH), Dana-Farber Cancer Institute (DFCI) and Massachusetts General Hospital (MGH). The UCLA affiliated institutions are members of the JCCC.They include UCLA Medical Center (UCLA), Olive View Medical Center (OVMC), VA Greater LA Healthcare System (VAGLAHS) and Cedars-Sinai Medical Center (CSMC).

Together, the Harvard/UCLA consortium can fully address the wide range of logistical and political issues involved in developing and implementing SPIN nationally. The choice of the particular Harvard and UCLA

affiliates participating in the consortium was made with consideration of the following needs posed by the complex SPIN RFA: development of a product, demonstration of feasibility for large scale implementation, and maximizing value to the research communities of the participating institutions. The Harvard group provides the primary informatics expertise needed at the core of this project. While UCLA members will be equal contributors, as a consortium we agreed that having the major development team located at one institution is preferable because it ensures a clear and unified vision of the product to be developed and an efficient centralized working setting to accomplish that goal. The combined academic centers and their affiliated hospitals provide a range of technical systems and organizational structures needed to develop and test a robust network. At the technical level, the 9 consortium members utilize 7 different pathology information systems. These include the major commercial vendors (CoPath, Sunquest, and Cerner), the VISTA system used throughout the VA networks, and several locally developed systems. At the organizational level, 2 major comprehensive cancer centers are represented, with distinct and complementary institutional organizations.

In contrast to the Harvard group , whose members are similar in structure and size, UCLA Medical Center is by far the major academic center of its group. This is a more typical situation for academic medical centers and affiliates nationwide, where a university hospital has one or more of affiliated VA, public, or private hospitals with varying degrees of academic activity and interaction among them. **We feel that all of this heterogeneity is a strength of the proposal, as it will account for issues that make our results generalizable**. Lastly, in developing SPIN, the effort expended must bring tangible value to the participating cancer centers and institutions. The DF/HCC and JCCC comprise two of the largest and most active cancer research groups in the nation, with joint participation in several major NCI-sponsored research programs ongoing or pending (lymphoma, soft tissue sarcoma, and prostate cancer) which involve members of Pathology and other departments. Making available the large and diverse collection of archived tissues to these research groups will facilitate cancer research in these areas.

C.1.1. On-going collaborative activities across the consortium

As members of the DF/HCC (P30-CA06516), the Harvard institutions have already participated in a number of successful collaborations including the establishment of 19 core facilities to support cancer research for the Center's 800 members. As an example relevant to CHIRPS, the Harvard-affiliated pathology departments have organized 7 technology- or disease-based Research Pathology Core Facilities across 4 of the Harvard hospitals with a Centralized Pathology Cores Administration located in the HMS Quadrangle academic pathology department. During the planning of the DF/HCC Core Facilities, two of the most common requests encountered were the need of researchers from all disciplines for access to well-characterized human tumor or normal tissue samples and access to histopathological interpretation of results and consultation with specialty pathologists. We responded to the latter request by establishing the DF/HCC Research Pathology Cores and Administration described above. However, the problems encountered in developing a common tissue repository were too large to overcome during the planning for the Cancer Center Support Grant (CCSG). The initial discussion on tissue-banking focused on the development of a centralized, fresh-frozen tissue repository. The logistics of collecting and storing samples that would meet the needs of researchers at multiple sites were possible though challenging, but they paled in comparison to the difficulties of obtaining common informed consent and the creation of an informatics network system that could permit queries across the Center's institutions. As a consequence, establishment of a tissue repository remained as a goal for the CCSG.

After the February NCI CCSG site visit, we began the second attempt to organize a Center-wide tissue repository. The effort began with the organization of an informatics network necessary to support a virtual specimen bank, which incidentally coincided, with the NCI objectives designed for SPIN. The DF/HCC effort is summarized in the following CHIRPS application. We will focus on addressing the issues of governance and operation, informed consent and establishing a web-based distributed informatics network query system for member institutions. The scope of CHIRPS will begin with retrospective indexing of archived specimens stored in the departments of pathology. At the completion of this phase of CHIRPS, we expect to have an informatics

network across the DF/HCC, which permits the query and access of specimens annotated with clinical information appropriate for the various levels of informed consent and IRB approvals obtained by investigators. The second phase of CHIRPS will be to expand the system for prospective tissue procurement and indexing, as proposed in the DF/HCC CCSG.

At UCLA, a centralized research pathology core was established in 1996-98 with ongoing support from the Jonsson Comprehensive Cancer Center support grant (P30CA19042). This facility, the UCLA Human Tissue Research Center (HTRC), is based in the UCLA Department of Pathology and Laboratory Medicine. In collaboration with the other affiliates, HTRC collects, stores and distributes human tissue specimens to affiliate investigators. The HTRC offers a full range of pathology related research services and has been included in virtually all JCCC associated research proposals.

C.2. Survey of pathology laboratory information systems

The 9 consortium institutions represent 7 different pathology information systems. In addition to customized programs, these systems include 3 from major commercial vendors (CoPath, Sunquest, and Cerner) as well as the VISTA system used by the VA hospital network. Although these heterogeneous data systems are a significant technical challenge, we believe that this arrangement is one of the major strengths of this application. This is because of our track record of integrating such systems and because we expect that any national extension of a distributed pathology informatics network system will have to manage this level of heterogeneity.

Table 1 below provides details of the volume of relevant pathology activities and specimen accrual across the consortium, including the percentage of surgical pathology records associated with retrievable paraffin blocks. It also serve as a snapshot of the LIS and HIS for the consortium. Appendix 2 includes the narratives of each participant's LIS and/or HIS, including where applicable, the access to cancer registry and other data. The extent to which queries can access clinical as well as pathology data is noted. Also included in Appendix 3 is an illustration of a composite data elements set in the consortium pathology information systems.

The consortium was able to take a different approach to the question of the completeness and validity of the data sets. The specific design of the existing institutional programs mandate complete data sets for the purpose of accessioning all locally generated surgical cases which by definition is the universe of cases with retrievable specimens. When BWH, MGH and DFCI came together to develop CRIS (see C.3.8 for description), a central data entry process confirmed that 95% of the source data was complete. The UCLA affiliates are currently completing this assessment.

	Surgical Path cases (1990-99)	% of cases with specimens	Online records since	Annual accrual rate	Hospital information systems	Pathology information systems
BID	175,000	93%	1995	45,000	Customized CCC	Modules from CCC
BWH	426,637	83%	1989	55,000	Customized BICS	Modules from BICS
CH	55,000	95%	1992	9,000	Oracle Database	Cerner AP module
MGH	517,000	85%	1976	63,000	Customized	CoPath
DFCI	3,000	100%	1997		Customized Oracle	CRIS & STIP
UCLA						
Univ.	245,170	95%		25,000	Customized	CoPath

Hosp.						
Olive View	~40,000	N/A	1990/1999	5,000	Compucare	Module
VA	75,333	N/A	1991	4,000	Vista/DHCP	Vista/DHCP
Cedars-Sinai	271,500	N/A	1987/1994	30,000	Customized ADS-Plus	Sunquest
Total cases	**1,768,494**					

Table 1. A summary of the number of surgical pathology cases online and percent of these cases with archived specimens, as well as their hospital and pathology information systems.

C.3. Relevant related work

The team members of the proposed consortium together have extensive experience in the development of distributed clinical and cryptographic systems as documented by the brief overview of the research, existing and pending, of the consortium's researchers.

C.3.1. Research Patient Data Registry (RPDR)

The Laboratory of Computer Science (LCS), allied with Partners Healthcare System (which includes BWH, MGH and other local hospitals), has developed the RPDR as a central clinical data repository for research. The RPDR combines data from various hospital legacy systems into one database. Researchers access this database through a real-time Web-based query tool that provides aggregate totals and additional patient identifying information with proper IRB approval. The RPDR brings massive amounts of clinical information to the researchers' fingertips but also ensures the security of patient data by controlling and auditing the distribution of patient data within the guidelines of the IRB. The RPDR currently maintains data on 1.8 million patients having 14 million encounters. The existing database occupies about 100 gigabytes of storage, and will grow to nearly a terabyte over the next 18 to 24 months as laboratory studies are added. Queries and their result sets are represented in XML in the software's middle tier. In December 1999, a RPDR was made operational as a pilot system. This effort parallels conceptually many of the same issues that will be addressed in the proposed CHIRPS initiative, including the use of XML for clinical data representation, Web-based query construction using large metadata databases, confidentiality models for large-scale databases, and the creation of a core dataset from disparate systems.[18]

C.3.2. XML-based clinical systems

The development of clinical information system architectures remains a key focus of LCS research. In recent years LCS has developed a number of active clinical systems based on XML with the highlights including 1) An XML representation of GLIF-based clinical guidelines, implemented for diabetes disease management[19] and 2) A complete electronic health record for Boston's Healthcare for the Homeless Program that uses an XML Portable Chart Format as the sole representation of data between the database and the application. [20, 21, 22] Of note, LCS is also the original source for MUMPS, the language/system on which many existing pathology information systems are hosted.

C.3.3. Distributed Multi-Institutional Querying

Since, 1994 members of this team at Children's Hospital have led efforts to share data across institutions with heterogeneous and disparate information systems, in real-time and over the public Internet.[32,34,35,36,37,42,43,44] This work, initiated under an NLM contract as the World Wide Web Electronic Medical Record System (W3-EMRS) project, has been influential in several ways. It has led/inspired several multi-institutional data-sharing projects[45] and has also led to the development of confidentiality policies for multi-institutional data exchange

published in the mainstream literature.[37] Another W3-EMRS related project is the BiliLIGHT system which provides virtual integration of birth data of mother and child from two birth hospitals (Brigham and Women's and Beth Israel) and incorporates these data into a real-time automated guideline for the management of jaundice.[39,44] This guideline is deployed at Children's Hospital at several sites and at selected affiliated practices over the Internet. An important part of the BiliLIGHT project was the authentication of multiple server machines across multiple competing institutions using a public key certificate system.[44]

C.3.4. Cryptographic identification systems

The PI of this proposed project has been a leader in the development of cryptographic health identification systems[40] and has subsequently been awarded an RO1 to work on the Health Information Identification and De-Identification Toolkit (HIIDIT).[33] HIIDIT is a distributed patient identifier system layered on top of public key cryptography. HIIDIT allows the designer of health information systems to select different judgements or trade-offs between competing desiderata for an identification system, such as who controls the creation and dissemination of identifiers, the extent to which the same identifier can be used for multiple purposes, the source of trust who certifies the identity of a patient or institution, the degree to which the identifier itself is kept secret, and the complexity of the resulting system of identification. That is, HIIDIT is not of itself a health identification system, but rather a **generator** of health identification systems. HIIDIT makes no commitment to a particular social policy, but it does define the major dimensions of the properties of any identification system and provides mechanisms for implementing various identification systems located in different loci within these dimensions. HIIDIT has been applied to multiple clinical domains including multi-center genomic databases.

C.3.5. Personally controlled clinical repositories

We have, with a group of colleagues, implemented a patient-controlled personal medical record system that follows our doctrines and supports our desiderata. Called PING (Personal Internetworked Notary and Guardian),[38] it was developed under the Federal Next Generation Internet Initiative. In PING, every record is essentially a collaborative document under patient control, enabling but not necessitating multiple data feeds from the collaborators in the patient's care. The PING record exists as a set of files available on a standard, publicly accessible web server (e.g. the patient's own area on America Online). Because the data are encrypted, only users with the appropriate roles can gain access to the information and the patient is given control over permissions granted to the various roles. Each PING records is stored as a set of XML (Extensible Markup Language)[47] files, which serve both as PING's repository of data and annotations about a patient and as a convenient format for communication of these data to other systems. XML is becoming a universal messaging format for the World Wide Web, hence it can be processed or displayed by programs and browsers on all common operating systems. Within the general XML standards, we are adopting document type definitions (DTD's) that correspond to medical communication standards.[48] To date, we support the HL7 DTD that permits PING to communicate with all current systems that use the HL7 standard.

C.3.6. Vocabulary and semantic classification systems

The Decision Systems Group (DSG) has had a long history of work in semantic structures and tools for medical concept management, including work with the Unified Medical Language System (UMLS) project of the National Library of Medicine.[49,50,51] A system known as Thenetsys was developed by J. Komorowski, E. Pattison-Gordon, et al, at the DSG, to edit and manage a semantically based thesaurus, and to browse it for updating and editing purposes, using a variety of visual paradigms. SAPHIRE, a retrieval system using concepts rather than terms, was initially developed in the DSG by W. Hersh et al. R. Greenes has worked for many years in structured reporting, and he and other faculty and fellows have applied this to encoding of clinical progress notes and radiology reporting.[52,53,54] R. Greenes and colleagues also developed a system for classification of radiology reports to enable targeted retrieval and feedback to radiologist of subsequent pathology diagnoses for the specific patients relevant to the body area examined.[55,56,57]

C.3.7. Natural Language Processing

Dr. Zeng developed several clinical applications with a Natural Language Processing (NLP) component while she was a doctoral student and fellow at Columbia Presbyterian Medical Center (CPMC), before joining the DSG as a faculty member. One major application was a decision support system that extracted and filtered relevant clinical data from CPMC electronic medical records. In that system, NLP technology was used to process radiology reports. She also worked with Dr. Friedman at CPMC in adapting an NLP system for ECG reports processing. At the DSG, Dr. Zeng has been working on various ontology projects and is knowledgeable about multiple vocabulary systems including SNOMED. She developed a new string-matching algorithm for mapping medical vocabularies to the Unified Medical Language System.

C.3.8. Clinical Research Information System (CRIS)

The Clinical Research Information System (CRIS) has been in place at the DFCI since 1997. CRIS is a relational database with two sources for the data: interfaces from current production ("legacy") systems, and direct data entry into CRIS. In addition to legacy data (appointment scheduling, ADT/Registration, Quality Control Center/Protocol Research, laboratory, and pharmacy) for all patients, CRIS also contains extensive clinical information on patients with selected diagnoses. Data on the baseline sociodemographics, clinical characteristics, treatment, and outcomes of women treated for breast cancer at the DFCI have been collected and stored in CRIS since 1997, and similar data for men treated with prostate cancer have been collected since 1998. The thoracic oncology and hematologic malignancies clinics at DFCI are scheduled to go live with CRIS in 2001. The linkages necessary to include the Massachusetts General Hospital in CRIS are currently being built, with rollout scheduled for late 2000. At that time, both legacy and clinical data from MGH will be available through CRIS. More detail about CRIS and the related Specimen Tracking Information Program (STIP) can be found in Appendix 4.

D. Research Plan

A successful approach to coordinating access to disparate tissue banks, or even disparate systems in general must account for variability in the local context. This variability takes many forms:

- Public perception – Public acceptance of making tissue specimens available nationally or internationally may vary.
- Local institutional policy – The institutions that establish and maintain these tissue banks may have reservations about making certain specimens available. Private tissue banks in particular may have valuable specimens but may not be able to share them in a comprehensive fashion.
- Technology – The systems that house the specimen information are certain to be heterogeneous. Even among the Harvard Medical Area sites there is no commonality in the approach to technology.
- Information completeness and representation – Different institutions will maintain different amounts of information on each specimen. Even when the same information is being kept on a specimen, the likelihood is that it will be represented in slightly different ways.

Consequently a practical solution to coordinating tissue banks must be simple to administer and administered locally. This assumption has a major impact on the system architecture that is described throughout the plan. The plan is organized with an initial section discussing governance of the project, and then chronologically by the three phases of the project as identified in the RFA. Within each of these phases we have organized our activities and subtasks within the themes of our three specific aims of data representation, consent and confidentiality, and distributed query architecture, so that the proposal can also be read thematically.

D.1. Governance

The local governance of the proposed project will take the form of three primary teams: a steering committee, a pathology/informatics advisory board, and an informatics task force. One of the goals of creating these teams is to delegate appropriately the key responsibilities of the project, and to avoid confusing policy and functionality

issues with technology implementation. For example, a specific assignment of the Steering Committee is to ensure the bi-directional flow of information between the NCI Coordinating Committee and the consortium. It is through this forum that the applicants will interface with and participate in the activities of the Coordinating Committee as described in "Terms and Conditions of Award". Additionally, chairs of each pathology department in the consortium have expressed their willingness to participate in the CHIRPS Steering Committee and have agreed in principle that specimens and information would be shared across the SPIN network. (See Section H 1 through H 6 for letters from each pathology department.) Whereas a function of the Pathology/Informatics Advisory Board is to enhance and foster the current and future collaboration among the experts in pathology-related informatics and the informatics community at large.

D.1.1. Steering Committee

This group will meet quarterly. Its role will be to assess the overall direction and progress of project. Christopher Fletcher, MD FRCPath, Professor and Director of Surgical Pathology (BWH) will chair the committee. Members will include the chairs of each pathology department of the consortium institution, a faculty in cancer outcomes, senior informaticians and chief information officers of each institution. They include:

Outcomes:	Jane Weeks, MD, DFCI
Pathology:	Douglas Anthony, MD, CH Pathology
	Jonathan Braun, MD, PhD, Chair, UCLA Pathology
	Robert Colvin, MD, Chair, MGH Pathology
	Harold Dvorak, MD, Chair, BID Pathology
	Peter Howley, MD, Chair, HMS Pathology
Informatics:	Octo Barnett, MD, MGH Informatics
	John Halamka, MD, PhD, BID Informatics
	Robert Greenes, MD, PhD, BWH Informatics
	Isaac Kohane, MD, PhD, CH Informatics
IT:	Joseph Bruno, Associate Dean of IT, HMS
	Michael McCoy, MD, Chief Information Officer, UCLA Medical Enterprise

D.1.2. Pathology/Informatics Advisory Board

This group will meet bi-monthly to review the direction of the informatics research and development as they apply to pathology informatics. It will be composed of informatics faculty and pathology faculty with expertise in informatics. Robert Greenes, MD, PhD, Professor of Radiology, and Director of the Decision Systems Group (BWH & HMS) will chair this committee. Members are listed below.

Henry Chueh, MD, MS, MGH Informatics	Isaac Kohane, MD, PhD, CH Informatics
Ulysses Balis, MD, MGH Pathology/Informatics	Thomas Drake, MD, UCLA Pathology/informatics
Stephen Black-Schaffer, MD, MGH Pathology	Daniel Valentino, PhD, UCLA Informatics

D.1.3. Informatics Task Force

This team will meet as needed. Its role will be to architect and implement the systems described in this proposal. The group is composed of informatics faculty, pathology faculty with expertise in informatics, and IT developers. Henry Chueh, MD, MS, Assistant Professor of Medicine, and Co-Director of the Laboratory of Computer Science (MGH) will lead the Task Force. Shawn Murphy, MD, PhD, MGH Informatics will lead the technical management of the project. Members are listed below.

Ulysses Balis, MD, MGH Pathology	Zuo-Feng Zhang, PhD, UCLA Tumor Registry
Bruce Beckwith, MD, BID Pathology	Alice Roberts, MD, UCLA Pathology
Aziz Boxwala, PhD, BWH Informatics	Honggang Shen, MD, UCLA Pathology
Barrett Goodspeed, MGH Pathology IT	Robert Trelease, PhD, UCLA Informatics

Gilad Kuperman, MD, PhD, BWH Informatics
Matthew Temple, DFCI IT
Shika Bose, MD, UCLA Pathology
Ralph Bowman, IS Manager, UCLA Pathology (Programmers)

D.1.4. Governance among the UCLA affiliate institutions

UCLA Medical Center, in contrast to the Harvard groups, is the major academic center. This is a typical situation for academic centers and affiliates nationwide, where a university hospital has one or more of affiliated VA, public, or private hospitals with varying degrees of academic activity and interaction among them. UCLA and its affiliates enjoy close interactions at research, clinical and teaching levels, though they are all economically independent. We propose to leverage this difference between the UCLA and Harvard groups, using the UCLA group as representative of the academic medical center affiliate arrangements around the country and a model for implementing and testing the proposed CHIRPS network in this setting.

In this model, UCLA will serve as the primary institution, and share responsibility with the affiliates in designing and implementing the network so that it serves the needs of the local affiliated group as well as the wider network. One particular aspect where this will be useful is in the investigation of approaches to incorporate data residing in settings other than Pathology information systems, such as clinical laboratory results and Tumor Registry at UCLA. In the latter example, we would first develop the process at UCLA, then roll it out to the affiliated institutions which share a common data system. This would provide an internal test of the ease of implementing CHIRPS on a wide scale as the RFA anticipates. With regard to the organizational and governance structure proposed, UCLA would represent the affiliates as well for the project overall, but would establish a comparable governance structure among the UCLA affiliates. UCLA and affiliates will therefore have a local governance/working structure that parallels in part that of the larger consortium. A steering committee made up of the faculty and IT representatives from each of the affiliates, and chaired by Dr. Braun will meet quarterly. There will not be a separate Informatics Advisory Board or Task Force at the local level.

D.2. ORGANIZATIONAL PHASE (YEARS 1 & 2):

D.2.1. Data representation: Defining data elements and data representation

A key building block of this proposal is to foster a consensus on what kind of pathology specimen data is useful and appropriate to share in the context of enabling specimen searches. Similar to the approach taken in the biomedical sciences, this effort can be thought of as a defining the key identification of a specimen as well as the more descriptive elements. We will describe this process as the "annotation" of a specimen. Consequently we are proposing the design of a SPIN Specimen Annotation (SSA).

D.2.1.1. Creating a SPIN Specimen Annotation (SSA) using Extensible Markup Language (XML)

Extensible Markup Language (XML) is emerging as the leading syntax for content-rich mark-up of information. Many Internet tools and software packages can already manage XML natively; for example, the latest version of Microsoft's Internet Explorer can read XML documents directly. We will investigate and identify how XML can best be used as a syntax for specimen annotation. The primary goal of this subtask will be to develop an XML specification for pathology specimen annotation. This specification will be described comprehensively using the new W3C standard XML Schema that is currently in working draft form. Though some aspects could be represented as an XML DTD, this form of markup does not provide sufficient expressivity in terms of datatype relationships. The specimen annotation specification will be able to scale from a minimal, least common denominator description of a specimen to a clinically content-rich description. Consequently many of the XML elements within the schema will be optional.

The basic elements of this XML definition will be derived from an examination of the fields being captured within the existing pathology information systems. Since these systems have been in operation for some time, they are an excellent source for common elements. In addition, we will be informed by related standards such

as the laboratory observation segments (OBX) in HL7. Data elements fall into three classes: 1) Required core elements that are common and exist in current systems, 2) A richer collection of optional elements that tend to be represented in many but not all systems, and 3) Extended elements describing additional clinical data that may or may not be captured in all existing systems. Through Dr. Balis, we plan to coordinate this activity with the data definition efforts of the College of American Pathologists.

The XML specimen annotation specification that emerges from this work will be comprehensive and inclusive with many optional and extended elements rather than minimal and exclusive. This will provide institutions an incentive to represent their data in this format, knowing that it can provide a bridge to future information systems. To the extent that a public HL7/XML specification is available that has information about pathology specimens, we will use it in the design of the specimen annotation in this project.

D.2.1.2. SPIN Specimen Annotation (SSA) definition

The design of the SSA will follow the hierarchical clustering of data and their natural relationships. For example, information that describes the patient who provided the sample will be organized together. This will make the SSA much easier to read and extend over time. The specimen annotation definition will be segmented into major sections for readability. Possible initial sections would be Source (where the specimen came from), Consent (a description of how specimen can be utilized), Clinical (supporting information such as clinical diagnosis, tumor stage, etc.), Specimen (descriptive information derived from processing and interpretation of the specimen itself), and Extended (additional information that does not fall into categories above, used for schema extensions as described below). In this proposed classification, the Clinical and Extended sections will likely be most expansive, based on the review of existing data. Shown below in Figure 1 is a partial snippet of an XML Schema describing a Patient as the Source of a Specimen. It emphasizes how data relationships can be defined quite specifically:

```
<xsd:schema xmlns:xsd="http://www.w3.org/1999/XMLSchema" xmlns:spin="SPIN.xsd"
targetNamespace="SPIN.xsd">
    <xsd:complexType name="SPECIMENANNOTATION" content="elementOnly">
            <xsd:element name="SOURCE" type="spin:SOURCEType"/>
             <!-- and so on ... -->
    </xsd:complexType>
    <xsd:complexType name="SOURCEType" content="empty">
            <xsd:element name="PATIENT" type="spin:PATIENTtype"/>
             <!-- and so on ... -->
    </xsd:complexType>
    <xsd:complexType name="PATIENTType" content="elementOnly">
            <xsd:element name="SEX" type="spin:SEXType" minOccurs=”0”/>
            <xsd:element name="AGESAMPLED" type="xsd:positiveInteger"/>
             <!-- and so on ... -->
    </xsd:complexType>
    <xsd:simpleType name="SEXType" base="xsd:string">
        <xsd:enumeration value="M"/>
        <xsd:enumeration value="F"/>
        <xsd:enumeration value="U"/>
    </xsd:simpleType>
     <!-- and so on ... -->
</xsd:schema>
```

Figure 1. A partial XML Schema for a SPIN Specimen Annotation, showing how complex data types can be defined and constrained. Note that SEX of a PATIENT can only be M, F, or U.

The proposed specimen annotation will take advantage of the extensibility of XML. Data modeling by consensus will form the basic data elements and relationships. Based on preliminary review, the required elements from the basic data set will likely be minimal and contain items such as Date of collection, Patient age, Patient Sex, Tissue type, and Clinical Pathological diagnoses. A specimen annotation containing only required elements would be considered valid. Other basic elements will be considered optional.

Additions and extensions to the initial XML Schema for SPIN Specimen Annotation (SSA) will occur in several ways. First, the consensus work will continue to refine and/or expand the basic data set. Second, researchers in specific domains can collaborate to extend the SSA to describe domain-specific detail. Third, local institutions can extend the SSA with local data that may not be applicable to other institutions. There will be two approaches to extend the SSA, both of which use the same XML Schema technique of deriving new data types from existing base data types:

1. Users can extend the definition of an existing datatype. For example, one could extend the PATIENT type with an additional element of HAIRCOLOR:

```
<xsd:complexType name="NEWPATIENTType" base="spin:PATIENTType" derivedBy="extension">
        <xsd:element name="HAIRCOLOR" type="xsd:string"/>
</xsd:complexType>
```

2. Users can extend the definition of the EXTENDED datatype that is intentionally empty initially, including as many additional complex data types with as much detail as needed.

In both of these cases, a local user can create a new XML Schema file with the same Namespace as the standard SPIN XML Schema for Specimen Annotation. A reference to this additional file would then need to be included in all specimen annotation files that used the data type extensions.

The advantages of this approach are clear. SPIN participants can modify the basic annotation for private needs without waiting for consensus from a governing body, and without altering the fundamental SPIN XML Schema. This will reduce any resistance to using a common annotation and prevent fragmentation of the core standard. As some extensions become widely accepted, they can be submitted for inclusion into the core Schema.

D.2.1.3. Packaging collections of specimen annotations as a query reply

Once a single specimen annotation is defined, there is an additional requirement to define how to package collections of specimen annotations. Packaging of individual annotations will be needed whenever one or more specimen annotations are transmitted; for instance, in the case of a query reply. In the model we propose, the standard result of a SPIN query will be a ***collection*** of specimen annotations. Metadata associated with this collection will also be included in the query result. Metadata includes information such as the query itself, the number of SPIN institutions participating in the query, and other statistics related to the resolution of the query from the SPIN network. The representation for a collection of specimens will deliberately not contain a hierarchy classifying the specimens themselves, since the annotation at the specimen level will have elements that can be sorted or classified easily using XSL transformations.

The flexibility of XML Schema allows us to create a separate schema definition for a collection of SPIN Specimen Annotations, but to reference the data types in the basic annotation schema. An XML instance document of a collection of annotations will therefore reference both schemata. The XML Schema for a collection will be created so that an XML document representing a collection can be visualized easily by transforming the collection with a single Extensible Style Sheet (XSL). We will define a default XSL stylesheet for a collection document, but additional styles can be created easily for local reports and other needs.

D.2.1.4. Vocabularies

For all of the formats discussed above, standard vocabularies will be used for those elements that have them. For example, the identification of laboratory values will be done using the LOINC identifier. Pathology diagnoses will be identified using SNOMED codes. Clinical diagnoses will be identified with ICD9 codes. In all cases the vocabulary coding scheme will be identified explicitly within the XML documents as an element or attribute. This allows the format to be as portable as possible and allows various coding schemes to be used in the future.

D.2.1.5. Coding natural language pathology reports

The goal of this subtask is to design natural language processing (NLP) tools that can construct an XML specimen annotation document from standard pathology reports in an automated fashion. In addition, an interface to review and edit converted reports will be designed. We will develop an information extraction (IE) application for pathology reports associated with the specimens in the databases. Since the purpose for our IE process is to index the specimens, the application will focus on extracting some important attributes of the specimens such as anatomical location and diagnosis and not attempt to understand the entire text. There are large corpora of pathology reports stored in the databases involved in this project. We will analyze these reports to develop a task-specific lexicon, named entity tagger, rules and patterns. The IE process will include four major components: preprocessing and local text analysis, extraction, template generation and review interface. In preprocessing and text analysis, the IE application will break the original text into sections, sentences and words. The words will be tagged with part-of-speech categories such as noun, verb, adjective and preposition. The application will recognize the named entities (e.g., body parts) specific to pathology specimen indexing with the help of the SNOMED vocabulary. Some light/partial syntactic parsing will also be applied to recognize phrases such as noun phrases. Extraction will perform domain specific pattern matching for simple facts and may produce additional facts by resolving co-reference and inference. For example, when a report says “The morphology of the tumor varies from area to area. In some places it shows a predominant glandular component.” we will resolve that the “it” refers to the tumor. All facts will be stored in a predefined entity structure that is consistent with the data model for specimen information described with XML Schema. In the template generation step, the facts will be arranged into a valid XML specimen annotation document using the XML Schema as a guide. Any exceptions will be written to an exception log. Examples of exceptions include missing required fields, repeated elements that are not defined as repeatable in the XML Schema, etc. The review interface will be used to browse and edit the generated XML specimen annotation alongside the exception log and the original narrative report.

D.2.1.6. Extraction of data from consortium systems

Beginning in year 2, we will ask consortium members to use the tools developed to create some example specimen annotations from their local institutional systems.

D.2.2. Confidentiality and consent: Developing a taxonomy of informed consent

The sharing of data involving potentially identifiable information is an important component of research, and great benefits accrue from it. Sharing among qualified researchers should be encouraged but only under authorized circumstances and under formalized processes that safeguard the confidentiality of identifiable information.[23] In nearly all areas of medical research, the use of identifiable information requires informed consent. Implementation of federal regulations, a straightforward process for many years, is now beset by ambiguity in a number of scientific contexts, particularly those involving human tissue and genetic analyses. As scientific opportunity has become richer and our analytical tools more powerful, the increasing complexity of study design and the possibility of secondary uses of data and specimens not anticipated at the time the specimens were collected have raised vexing questions about the nature of a proper informed consent.[27,28]

We recognize that issues of informed consent for human cell repositories are controversial with no uniform approach as yet agreed upon by all governing bodies.[24] In this application, we make every attempt to conform

to the OPRR guidelines. **By creating a sufficiently versatile consenting taxonomy in the specimen annotations, we intend to be able to accommodate future policy decisions arrived at the federal and/or local level**. Our priority is the protection of patient confidentiality and the rights of those individuals to privacy. In addition, we realize that patients have the rights to direct the manner in which their specimens are managed.

D.2.2.1. An initial model for a consent taxonomy

For the tissue repositories in our consortium, the collection of specimens has been performed under differing conditions. It is expected that this is the typical case throughout the consortium. As it is currently not possible to detail this on a 'repository specific' basis we will attempt to create a basic taxonomy of informed consent circumstances and how these might be managed in the CHIRPS system. The following are the levels of security, which exist for specimen repositories, and our initial proposal for the allowable research to be conducted under these circumstances. We anticipate that the taxonomy will evolve substantially during the first two years of the CHIRPS project. The following sections describe the characteristics of these levels of consent and their implication with regard to specimen acquisition.

Level 1 – Specimens without specific informed consent

This often pertains to specimens collected under routine clinical consent only. Archival material from these specimens that is stored material for clinical documentation, reference and management purposes in pathology departments is necessarily stored with identifiers. Patient material collected under routine clinical consent in excess of that used and retained for the preceding purposes is usually discarded, or, when retained for teaching or research purposes, is designated "excess patient material". This material may be banked or otherwise utilized under guidelines approved by the institution in which it is collected, and depending on the routine clinical consented use, is generally anonymized.

In this case it will be important for IRB review of each request to determine level of risk to patients and therefore whether or not informed consent is necessary. The following guidelines are given, although determination of definitions of risk, and where a given study falls in this regard, lie at the level of the institutional IRB's:

1. Minimal risk studies - If there is a minimal risk to the individual participant, disclosure of data may be approved by an IRB without requiring the investigator to re-contact the participants. For example, somatic mutations in a group of cancers from patients who have since deceased.
2. Moderate risk studies – Attempts should be made to consent individuals. Only under circumstances where reconsent is impossible (prolonged passage of time, death of patient, number of specimens >1000) should waiver of informed consent be considered and then only with institutional IRB approval.
3. High risk studies – for example, germline DNA testing or highly sensitive information being obtained e.g. criminal record linked to blood type. A new informed consent for disclosure of the data **must** be obtained from the research participant, or the study will not be allowed. In this case the investigator requesting the data should obtain IRB approval to contact the prospective participants, explain the proposed study and seek their informed consent. Additional general guidelines are described in Appendix 5.

Level 2 – Specimens with research-specific informed consent

At this level consent has been obtained to ask specific types of research questions, with some restrictions. Investigators will have access to these specimens provided that the type of research being conducted complies with the intent of the original informed consent. If research described in the proposal is closely related to the specific research for which the informed consent was obtained, the Bank's institutional IRB must sanction use of specimens for this purpose.

Level 3 – Specimens with blanket informed consent

Here consent has been obtained at the time of tissue retrieval to use specimens for research purposes, not otherwise specified, with no restrictions. In this case, sharing of data with other investigators for IRB-approved research is covered in the original informed consent under which the data were collected. Specimens and data,

stripped of identifiers, will be given to researchers upon request, provided that the submitted proposal complies with guidelines stated in Appendix 5.

D.2.2.2. Implications of the consent taxonomy for the specimen annotation

A section detailing the level of consent will be included in the specimen annotation. If the only information required was the level of consent, a single XML element could be used for this purpose. However, additional information such as the categories of research allowed for Level 2 specimens is clearly useful. In addition, as our understanding of consent issues evolves through national debate and discussion, details at each level of consent are likely to emerge. Consequently the XML representation of the consent taxonomy will take the form of explicitly identifying the level of consent in the XML Schema as a complex datatype. In addition, since multiple taxonomies may emerge to classify consent, the taxonomy used will also be identified explicitly. This approach allows the development of the consent taxonomy to be loosely coupled to the development of the specimen annotation.

D.2.3. Implementation: Design of distributed architecture

D.2.3.1. A peer-to-peer SPIN network

Many large-scale repositories are centralized databases that are able to avoid the complexities of a distributed model. A national tissue bank suggests a distributed model for several reasons. First, institutions will need to continue to maintain source tissue banks and databases locally. Therefore, a centralized model would require the maintenance of both local databases and sophisticated synchronization to a central database – not a practical solution. Second, the idea of maintaining a national database that contains all specimen data may be contrary to public policy. We assume, then, that a national specimen databank will need to be composed of distributed databases, each located at the site that generated the specimens. These databases need to be connected through a network to form a virtual database.

A network model to connect these databases can be constructed using a hierarchical or a peer-to-peer topology. In the hierarchical approach, one site would be responsible for managing information about access to other sites' databases. Hierarchical models have the benefit of potentially simplified site discovery and query access, since this master site can maintain some information about all other sites. However, a distinct disadvantage of the hierarchical model is that it requires more governance and effort to establish and maintain. It also begins to create a centralized repository of information, however minimal. In addition, if the "master" site fails for whatever reason, the network would come to a halt.

An alternative approach that we support in this proposal is to design a non-hierarchical, peer-to-peer network for indexing and locating tissue specimens. In this model all institutions that want to participate are peers. **This architecture leverages the decentralized model of the Internet**. Any number of SPIN sites could fail or be offline and the network would remain viable, much the way that many nodes on the Internet can fail without affecting the Internet as a whole. In this model the virtual SPIN network for publishing tissue specimens is established through the definition of an Internet-based SPIN connection protocol. Any institution that wishes to participate can obtain the freely available software that uses the SPIN protocol to establish a presence on the SPIN network. The choice of participating in the SPIN network remains a local decision to become a SPIN peer.

D.2.3.2. The proposed SPIN connection protocol

The model proposed here for the SPIN connection protocol will be similar conceptually to the protocol used for the Gnutella network.[1] This protocol will be used to communicate between SPIN servers at different sites. The protocol will use basic TCP/IP socket connections over a designated port. The connection protocol is intentionally simple. Except for a simple handshake to establish a connection, all other communication will consist of a single data packet consisting of a header, a message type, and data associated with the message. There will be three primary message types: a one-way "Announcement" message and a "Query/Reply"

message pair. Announcement messages will be broadcast by a SPIN site to identify itself to other SPIN sites in a "friend of a friend" approach. Query/Reply messages will be used to send specimen queries and receive replies. Connections are terminated immediately after the message packet is sent. The basic message types, data and routing rules associated with them are summarized below:

MESSAGE TYPE	DATA (XML)	ROUTING
Announcement	Number of specimens online for querying	Send to all connected sites Do not route duplicates
Query	Specimen query	Send to all connected sites Do not route duplicates
Reply	Collection of specimen annotations	Return to sender of Query

Using this protocol, a SPIN compliant server will do the following:

1. Upon starting, prime an internal list with one or more active SPIN sites, connect to these sites, and send out an Announcement message over all successful connections (the very first site does not do this, of course).
2. When an Announcement message is received, forward the message to all connected sites and add the IP address and port of the announcing site to the internal list.
3. When a Query message is received, make note of the message ID and sender, forward the message to all connected sites, perform the query locally, and send a Reply message back to the sender of the Query.
4. When a Reply message is received, match the Reply with the previously noted Query (using the message ID) and send the Reply back to the associated sender.

A collection of servers on the Internet running software that supports the described SPIN connection protocol constitutes the SPIN network. Queries are constructed separately and passed into this SPIN network for distributed processing. This separation of query construction and query processing is described in the next section. Overall, this model allows many different parties to develop SPIN compliant server and query tools software. We will develop one form of server software and query tools as part of the CHIRPS initiative.

D.2.3.3. CHIRPS clients and servers

In an ideal peer-based network, all clients are potential servers, and vice versa. However, the software required to establish a CHIRPS server, while simple to install, may constitute too significant a barrier for an investigator who wants to simply locate a specimen but is not ready to publish a specimen repository. In addition, since one subgoal is to ensure that query access is available through a standard Web browser, the software that constitutes a CHIRPS server will be distinct from the software used to drive the user interface for query construction. In the proposed model, CHIRPS servers will act as both client and server to each other, and contain embedded software to respond as a Web server to a standard Web-browser client and user interface. The servers will need to communicate with a centrally maintained authentication site to establish an island of trust among the collection of server sites. X.509 digital certificate technology and SSL will be used to securely authenticate servers. The servers can then verify the identity of peer servers, allowing the exclusion of non-registered servers. This model allows for multiple islands of trust, as well as collections of servers that are not secured because they are located on a secure intranet, for instance.

The following schematic provides an overview of this model:

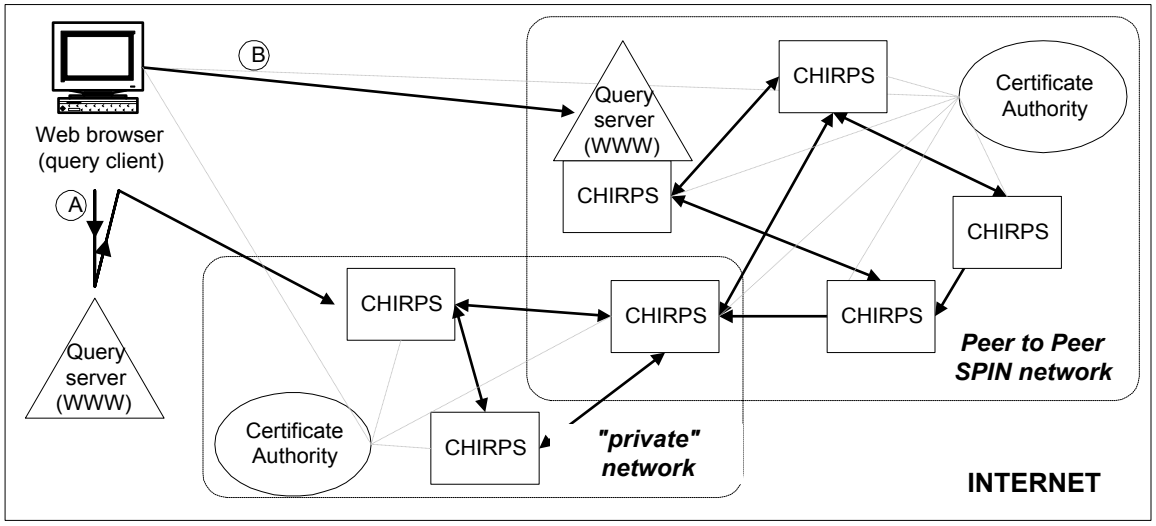

Figure 2. Here the SPIN network is represented as a single island of trust on the Internet, implemented by a peer-to-peer network of CHIRPS servers. A query can be composed from any standard Web browser pointed to (A) a SPIN query construction Website that does not host a CHIRPS server, or to (B) a Website hosted by a CHIRPS server node.

D.2.3.4. Selection of local specimen annotation storage

We have already proposed that all information related to specimens remain stored locally at each institution. The existing native storage systems for this information are quite varied, as reflected by the different systems used by each institution in our consortium. There are two basic approaches to exposing these local databases to the SPIN servers and network. One approach is to require SPIN participants to create a direct interface to their local system that can respond to queries and return a collection of specimen annotations. This model is "tightly coupled" in that very specific local programming will have to be performed to allow systems to respond to queries directly. In this scenario a single SPIN query by an investigator would invoke numerous native queries running on heterogeneous pathology systems. The primary advantage of this model is that it takes advantage of the existing system. However, it suffers from a number of disadvantages, including:

- Developing an "adapter" that performs direct queries against a particular pathology system/database is not trivial, and constitutes a significant barrier to participation.
- With heterogeneous systems, query performance may vary dramatically from institution to institution.
- Invoking queries directly against an operational, production pathology system may cause performance problems to local applications such as those used in the pathology laboratories.
- The tight coupling of existing pathology systems into the SPIN network in this model means that any major system upgrades or changes in the institutional platform will likely "break" the link.

For all of these reasons, an alternative approach for providing local specimen annotation storage is proposed that does not involve direct queries to local systems. Instead, a "publishing" model is proposed whereby tools will be provided so that participating sites can easily publish all or parts of their local database into a separate, defined repository[25] (See Figure 3).

A number of advantages accrue with this approach:

- The local CHIRPS server storage provides a kind of a "DMZ (de-militarized zone)" that isolates the SPIN network from local production systems. This reduces the chance of errant queries or security problems affecting local production systems.
- Query performance can be made much more predictable.
- Tools to manage the local annotation storage can be provided as part of the CHIRPS server software, since the local storage model is now consistent.

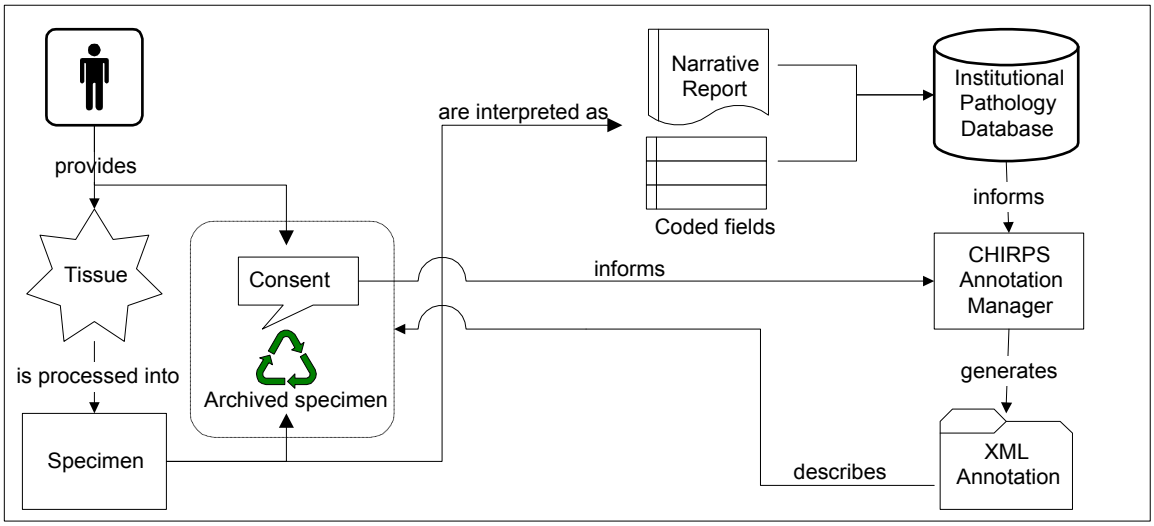

Figure 3. A schematic of information flow, showing the Annotation Manager tool that is provided as part of CHIRPS software to establish a local XML annotation repository.

To keep the CHIRPS server software simple and easy to implement, we propose that XML specimen annotation documents will be kept in file directories, rather than a database. Consequently the local annotation storage system is simply the file system of operating system that the CHIRPS server is hosted on. The CHIRPS software will support an internal database to maintain high performance searching for the required data elements. Optional data elements would be searched through file text searches. Although this places a burden on the CHIRPS server software to have appropriate techniques for searching these directory-based files, it simplifies the implementation of a new SPIN site in a number of ways:

- Institutions can easily manage specimen data by using standard file management techniques.
- There is no need for extensive third-party database software.
- Directory organization can be used to easily establish various categories of specimens, some of which could require additional authentication. Basic directory security is routinely provided by all operating systems.
- Basic file and directory services are available across operating systems on all platforms.

On the CHIRPS server where CHIRPS software is installed, a set of file directories will contain XML specimen annotations. This server should exist behind a firewall, with only the TCP ports needed for CHIRPS peer-to-peer connections exposed to the Internet. Note that the production pathology systems also remain safely behind the firewall, and do not even need to be accessible through any proxies in our proposed model. CHIRPS tools help the institution convert existing data from pathology systems into XML specimen annotations in the file directories on the CHIRPS server. More details on the implementation are available in Section D.3.3.

Problems that we have encountered in the past using this approach include:

1) Changing primary keys. If a new primary key is established in the production pathology database, all the files need to be reread and the key updated. While this can be a lengthy process, it is basically a re-export. Primary key changes happen infrequently because of the effort involved on just the local system.
2) Adding new data to all the existing files, for example, if one wants to update ICD9 to ICD10 codes in all the files. This can be managed with appropriate software tools.
3) Maintaining synchronization with the internal database. Since the file system provides great freedom in making changes to the files, it is relatively easy to alter files so that they are invalid or inconsistent with the local production database. This can be managed with tools that validate the local annotation store.

D.2.3.5. Query representation and construction

The query syntax will be Extensible Query Language (XQL), which is also represented as XML. This will simplify the extraction of data from the XML annotations. The user interface for queries will be designed in HTML for standard Web browsers. The approach we propose is to construct the queries outside of the SPIN network using standard Web server application techniques, and then to send the query into the SPIN network for resolution.

D.3. COMPONENT SELECTION, DEVELOPMENT & IMPLEMENTATION PHASE (YEARS 2-3.5):

D.3.1. Data representation: Distributing and processing queries

D.3.1.1. A distributed query mechanism

The model for query distribution will be based on the virtual SPIN peer network. Using the model established by the Gnutella Network, every institution running SPIN software on the Internet will be discoverable by each other (see D.3.3 for details). Once a query is received by the SPIN server, it will be sent to all other reachable SPIN servers. They in turn will forward it to other SPIN servers, and so on.

D.3.1.2. Composing SPIN queries: SPIN clients

SPIN queries will be represented fundamentally as an XML document in XQL. Functionally, these queries can be composed on a client web browser in communication with a Web server that hosts a query tool. The resulting XML document is then sent to a CHIRPS server using an HTTP POST. Metadata for the construction of the queries will also be served from the site supporting the query tool. The advantage of defining explicit query syntax such as XQL/XML is that it will support future research into the independent development of increasingly sophisticated query clients. The XML representation of the query will enter a CHIRPS node where it will begin its distribution through the peer-to-peer network. The results on the query will then be served back the query tool site from the initial CHIRPS server. Query tools can also reside and be maintained at a CHIRPS server site, allowing for specialized queries to be constructed using specialized user interfaces. The use of an XML intermediate representation will allow advances in query construction and visualization to proceed independently, while at the same time the SPIN network can support any number of SPIN compliant query tools.

D.3.1.3. Processing queries

The process of how the XML annotations at each site repository are queried is controlled by the CHIRPS Query Engine. This process is described below. However, each site will have control over what results it returns when a complete specimen annotation is requested. Although all required elements must be included (or the annotation would be considered not valid), all other elements can be included or excluded based on the discretion of the host of the site. An XSL template will be available for each site to customize to achieve this filtering.

D.3.2. Confidentiality and consent: Enforcing security policies with technology

D.3.2.1. Authentication of a SPIN client and staging the access to information

A CHIRPS server can be established independently by simply implementing the freely provided software. However, to cooperate with other servers in a network, a digital certificate to establish server authentication will be required. Query messages may then come from a user query tool or be routed from another CHIRPS server. Each individual CHIRPS server may choose to limit IP addresses from which queries may be served or require passwords just in order to limit volume. However, the queries themselves are then propagated through the SPIN network in staged security modes. Stage 1 queries that require no security provide only limited, aggregated information contained within the CHIRPS server's internal database. The information is guaranteed to

be de-identified by the CHIRPS software that loads the data from the XML files into the internal database. This is essentially open to any query tool that can post a valid query into the SPIN network. A Stage 2 query would be one that returns individual specimen data that is limited to the required data elements, while Stage 3 queries allow for a result set to include comprehensive specimen annotations. Stage 2 and 3 queries would require use of the digital certificate authentication. In this mode the user would be required to present authenticating credentials with his query to obtain more detailed information from the XML specimen annotations. An example of a query not requiring authenticating credentials is: "How many specimens of breast carcinoma exist on the SPIN network?" An example of a query requiring credentials is: "Show me all specimens of Grade III breast carcinoma for women under the age of 35, and in the result set include age, institution, race, and clinical history inclusive of medications and other diagnoses".

A researcher would use a typical web browser to conduct their query of the CHIRPS system. A typical query might begin with a search for all specimens, which have a tissue type of "adrenal" and a diagnosis of "pheochromocytoma". For this type of query, no informed consent is needed for the investigator. A list of the number of matches would be returned. With appropriate authentication (which may require IRB approvals), the investigator could then iteratively refine the search by specifying various criteria, such as sex, age range, date of diagnosis, etc. Once the query has been suitably refined, the researcher will be able to view the list of specimens and select individual specimens to request electronically. The requisition should be in the form of a proposal with detailed description of how the specimens would be used, as well as certification of appropriate IRB approvals. The encrypted requisition would be forwarded to the various institutions, which houses the specimens of interest. If the requisition were deemed bioethically and scientifically meritorious by the governing board of the institution(s) housing the specimens, and if the said institution approves, an agreement could be entered by the parties.

D.3.2.2. Process for obtaining the specimen

A query that is performed in the "secured" mode would allow for full identification of the site at which the sample resides, as well as contact people available at the site. Information about the sample could then be used to obtain the sample from the site. There is no direct link between sample numbers and data retrieved in the query. The query would need to be rerun internally at the site to link to specific sample numbers. Each site would then have full control over policy and etiquette regarding the actual delivery of samples. These policies and procedures would be hyperlinked from all CHIRPS responses.

Types of tissue banks and guidelines for access:

Across the Harvard/UCLA consortium, there are over 100 specimen repositories, categorized into two types—private and public. "Private" banks are developed by investigators interested in a particular tissue-type or disease process. "Public" banks are developed by clinical departments and institutions. For the scope of CHIRPS, we propose to create an informatics model system to allow query of the archived, annotated specimen banks established by the various departments of pathology in the decade of the 1990s. In general, these archived specimens are residual or excess patient tissues in the form of formalin-fixed paraffin blocks generated from surgical pathology cases, after appropriate samples are taken for clinical diagnosis. Excess or residual tissue is commonly defined as tissue removed at the time of surgery, which is not needed to establish a clinical diagnosis. The CHIRPS informatics model system will be built to allow participation of private banks.

As described above, investigators requesting access to specimens should formulate the requisition in the form of a proposal with detailed description of how the specimens would be used, as well as certification of appropriate IRB approvals. This will include use of a CHIRPS-sanctioned standardized requisition form. The requisition should be de-identified, encrypted and then forwarded to the various institutions, which houses the specimens of interest. If the requisition were deemed bioethically and scientifically meritorious by the Governing and Operating Committee of the institution(s) housing the specimens, and if the said institution approves, a binding agreement could be entered by the parties.

The Harvard institutions already have an existing Pathology Cores Oversight Committee governing the activities of the DF/HCC Research Pathology Cores. This Committee consists of al the pathology department chairs and a representative from the Harvard School of Public Health. To establish the CHIRPS Governing and Operating Committee, we will invite an expert in informed consent and a bioethics consultant to join the group. This CHIRPS Governing and Operating Committee will be responsible for ensuring that the banks meet all existing institutional, legal and ethical standards, as well as monitor compliance through the mechanism of annual review of IRB protocols in accordance to the established institutional guidelines. The Governing and Operating Committee will also review requests for access as well as establish appropriate users fees for specimen retrieval, histologic evaluation and certification so as to ensure continual distribution of specimens and that the banks could remain fiscally sound.

D.3.3. Implementation: CHIRPS software and the SPIN network

D.3.3.1. System and security architecture for CHIRPS

The CHIRPS system architecture is built around three principle components, the peer to peer CHIRPS servers, the query tool server(s), and a certificate server. In order to be part of a CHIRPS network, one needs to obtain a digital certificate from the certificate server. A local certificate server can also be created to maintain private CHIRPS networks or sub-networks. In the above diagram, the circles represent the certificate servers and the dotted lines are connections to the certificate servers. When data is transmitted from one CHIRPS server to another, the recipient checks the certificate to ensure a registered CHIRPS server is sending the data. SSL and X.509 protocols will be used to support this feature. This model ensures that a particular collection of CHIRPS servers can trust each other, forming an "island of trust".

When a Stage 2 or 3 query is made (see above for descriptions of confidentiality levels), a certificate will also need to be provided by the internet browser client. This is over and above any authentication the individual CHIRPS server may require to control access to query initiation on their server. The clients certificate can then be checked to ensure it is a registered client. Furthermore, each CHIRPS server can keep a list of acceptable client certificates for Stage 2 and 3 queries. This will help a CHIRPS servers both control the volume of queries on their server (because these queries may utilize considerable server resources) and control access to this data.

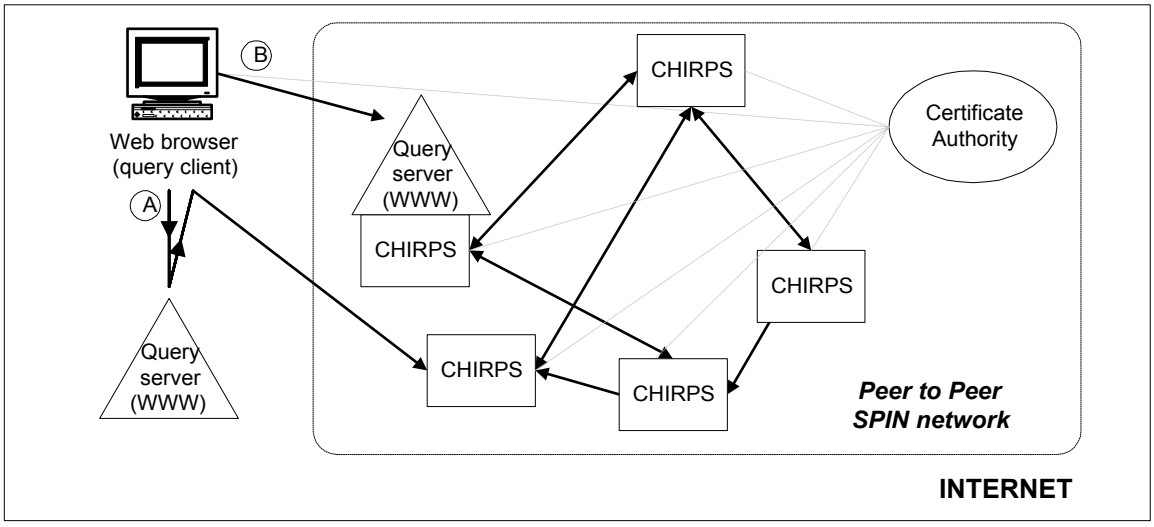

Figure 4. An overview of the proposed SPIN network using trusted CHIRPS peer-to-peer servers.

D.3.3.2. Query construction and processing architecture

A query will be initiated from the query tool server. As part of this proposal, software to establish a default SPIN query tool server will be implemented. However, other query tool servers that support specialized query

construction can be developed. These are shown as triangles in the diagram above. The query tool will take the user through a session on their Internet browser using typical metadata tables that will be found on the query tool servers. We have constructed such query tools in the past and their technical problems are well understood by the members of this consortium. Once the query has been constructed by a visual interaction in the Web browser, the query tool software will construct an XML representation of the query (as XQL). A Stage 1 query can then be submitted to any CHIRPS server that is willing to accept the initiation of the query through the network. Stage 2 and 3 queries will need to provide their digital certificates. Once the XQL query arrives on a hosting CHIRPS server, that server will not only begin processing the query locally, it will also forward the query to its known peers. The query will be passed peer to peer as has been described above, so that all of the CHIRPS servers active on the Internet have a chance to perform the query. In some cases a CHIRPS server will not run the query because it does not recognize the digital certificate associated with the source CHIRPS server. However, it will still forward the query. The query results are then routed back and eventually accumulate back on the initiating CHIRPS server. The results are then fed back to the query tool server as a single XML document representing a collection of specimen annotations located on the network that satisfy the query. The query tool server then has the responsibility to manage the presentation of these results to the user's Web browser.

D.3.3.3. Local XML repository and query engine

The fundamental unit of the CHIRPS system is the peer to peer CHIRPS server. These are shown as squares in the diagram above. The CHIRPS server will communicate on the Internet through standard TCP/IP/SSL protocols. The repository of pathology data exists on this server. A directory in the local file system of the server will contain the XML files that represent pathology specimen annotations. The grain of the files will most likely be one file per specimen. The files will be created using the CHIRPS Annotation Manager, a toolkit for importing data extracts from local databases. Fundamentally the toolkit will work by requiring at least a table or view that is organized as one specimen for each row. The local database would then de-normalize linked tables so that a logical view of the data containing one long row for each specimen will be obtained. Each row is then placed into a separate XML file. For simple database schema, this process will be performed using SPIN supplied software. This software will perform in a similar fashion to commonly available commercial software such as Microsoft Access, which performs the inverse (normalizing) transformation. For complex database schema, the XML in the files will be created by a process equivalent to that uses to represent RIM data in HL7 messages. In fact, it is likely that the software that is used with large, complex databases will be a commercial tool to generate HL7 Version 3 messages that are then transformed into XML using our toolkit.

Once a repository of XML specimen annotations has been created, it is queried by the CHIRPS Query Engine. It is currently envisioned that all software components supported by our consortium will be written in the Java software language. The plans are to use the Sun Microsystems JDK as the interpreter, to best achieve a platform independent set of tools. The CHIRPS Query Engine will have a very basic internal database that will index the core, required elements that should be present in all of the XML documents. This database software will be developed as part of the Query Engine so third-party database software will not be needed (eliminates some potential platform dependence issues). Software to synchronize the files with the database will be part of the Annotation Manager. Keys would include such items as diagnosis codes, dates, and other minimal information.

Based on the above described method of organizing the data, queries would take on essentially two forms: highly structured queries using the required fields from standard metadata, and more loosely structures queries that would most likely require knowledge about that particular domain of pathology. The first type of query (a Stage 1 query) would use the high-performance internal database at each site only, while the second type of query (a Stage 2 or 3 query) would need access to the individual XML annotation files. All queries beyond Stage 1 would exploit those data elements that will vary from specimen to specimen; for example, the ELEMENT <THICKNESS> may only exist in annotations of skin specimens. In the circle of skin pathologists this may be an obvious field to include. It could then be included by general agreement or arbitrarily. This type of query would necessarily take longer to run (since it needs to look through individual XML files), and also

poses a greater security risk to the hosting site. The CHIRPS Query Engine will filter and bin the data, so that no patient can be identified through the Stage 1 queries. However, this degree of security will be a much more difficult, if not impossible, to achieve with the Stage 2 and 3 queries given the open structure of the XML files. For those types of queries, security is managed through the process of authentication, as well as the ability to filter out specific data elements from the annotation before replying with the query result.

The other pieces of the CHIRPS server software include the Certificate Server and the Query Distribution Application. As discussed above, the Connection Manager will be modeled after the architecture of the public domain, open source code project Gnutella and will not be further elaborated upon here. The Certificate Server can be any commercial product that follows X.509 standards.

D.3.3.4. CHIRPS software components

The CHIRPS software will be designed as collection of interrelated components written in the Java language to be multi-platform.

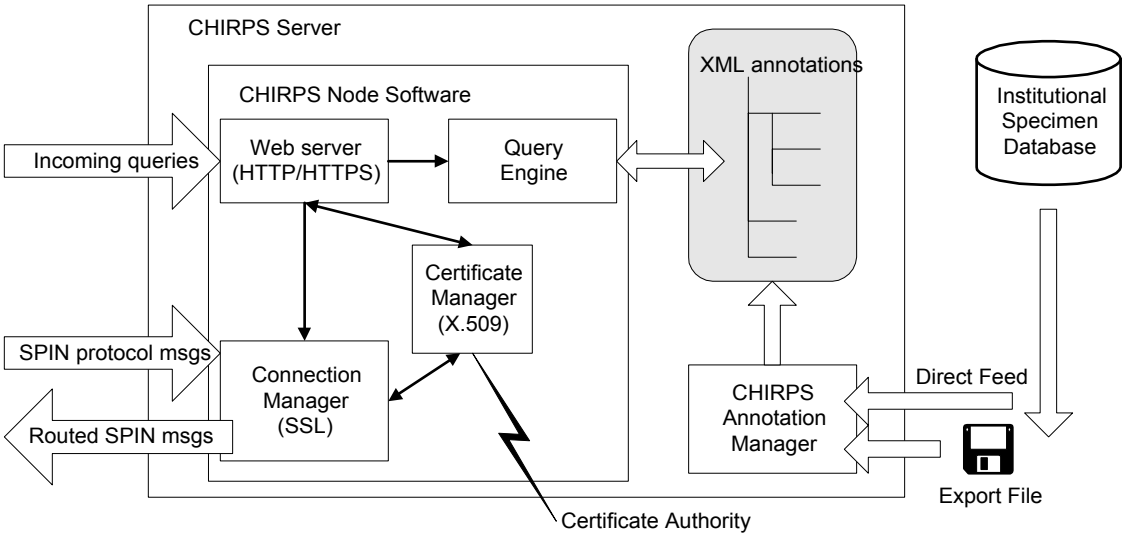

Figure 5. Diagram of the CHIRPS server and associated software components

Connection Manager – This is the core component that manages all communication between SPIN servers over the Internet using the SPIN Connection Protocol.

Query Engine – This component accepts an XQL query from the internal Web server and processes the query against the XML annotation store. It uses an internal index of the minimal required data elements of the Specimen Annotation to enhance performance. Once a subset of files is identified this way, individual XML files are then queried in much grater detail.

Web Server – This is a minimal implementation of a Web server that can support HTTPS. It will accept an incoming query from any query construction application as an HTTP POST of an XQL document. It will also accept periodic requests for query results related to an XQL query instance.

Certificate Manager – This module will manage all aspects of verifying X.509 certificates, communicating with Certificate servers, and signing outgoing messages routed by the Connection Manager.

Annotation Manager – This is a suite of software tools that includes all components needed to process and import local institutional data into the XML Annotation store. The NLP tools are part of this suite. This tool set

includes applications that can establish an interface for real-time direct feeds (HL7) or batch imports from the local systems.

D.3.3.5. Implementing the connection protocol

TCP/IP and SSL will be utilized to implement the CHIRPS connection protocol. The Secure Sockets Layer (SSL) layer of HTTPS can support full 128-bit encryption if sensitive information needs to be transmitted. When a CHIRPS server is started, it will join the SPIN network by announcing itself to at least one other CHIRPS server on the network (the first CHIRPS server to start will not have anyone to announce to). A SPIN Web page will post a list of known CHIRPS servers for informational purposes. Each CHIRPS server will maintain locally a list of other CHIRPS servers on the SPIN network. This will occur automatically, with the only assumption that the list is primed with the IP address of at least one other active CHIRPS server.

D.4. TESTING AND VALIDATION PHASE (YEARS 3.5-5)

Prior to this phase, we will construct a detailed plan for testing and validation, based on the results of the previous two phases. The plan will have the major components described below.

D.4.1. Demonstrate that all components work at each institution

We will design a test workplan that exercises all aspects of the proposed architecture, from the submission of new basic data elements for the SPIN Specimen Annotation Schema, to processing and importing narrative pathology reports into XML. Based on feedback from consortium sites, we will:

- Correct software and hardware errors and improve performance of system
- Develop operation and repair protocols for the system
- Create documentation for how to participate in the SPIN network, and for how to use the CHIRPS software tools

D.4.2. Performance testing

D.4.2.1. Testing SPIN server discovery

We will recruit at least one site in every state to establish an instance of a CHIRPS server as a mock SPIN node. No specimen data will exist as part of this test. The test will consist of both timing and completeness in terms of discovery of all nodes on the SPIN network. Each CHIRPS server will be primed with several randomly chosen other CHIRPS server IP addresses (identity selections will be eliminated) as well as the consortia's SPIN address.

D.4.2.2. Testing SPIN query performance

In this subtask a list of queries will be developed by the Steering Committee. These will be transformed into XQL by our team, and sent into one of our consortium's CHIRPS servers for resolution. Timing and assessment of query results will be the outcome measures.

D.4.3. Usability testing

D.4.3.1. The default Web browser based SPIN query composer will be provided to no less than fifteen members of the research community. Each will be asked to perform several tasks, including:

1. Perform queries from a canned list of narrative query descriptions, using the Web query composer.
2. Describe and perform self-initiated queries using the Web query composer.
3. Complete a questionnaire rating the accuracy and effectiveness of results returned for each of the canned queries.
4. Complete a formative evaluation instrument designed to address usability, perceived speed, and other qualitative factors.

D.4.4. Follow-on development and dissemination

Corrections to the software will be made based on results from the performance and usability testing. Repeat testing and targeted evaluation will be performed only as necessary. An initial version of the CHIRPS software will be made available on a public SPIN Website along with documentation.

E. Human Subjects

We will use clinical information obtained from patients in the development of CHIRPS. Applications have been submitted to the IRBs of Harvard Medical School and the University of California at Los Angeles, and approval is pending.

F. Vertebrate Animals

No.

G. Literature Cited

1. Gnutella Network. <http://gnutella.wego.com/>
2. Lubeck, D. P., Litwin, M. S., Henning, J. M. et al: The CaPSURE™ database: a methodology for clinical practice and research in prostate cancer. CaPSURE™ Research Panel, Cancer of the Prostate Strategic Urologic Research Endeavor. Urology, 48:773, 1996
3. Van der Spek-Keijser LM, Van der Rhee HJ, Toth G, et al. Site, histological type, and thickness of primary cutaneous malignant melanoma in western Netherlands since 1980. British Journal of Dermatology. 136(4):565-571, April 1997
4. National Cancer Institute, Cancer Diagnosis Program. <http://www-cdp.ims.nci.nih.gov/resources.html>
5. Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD. The cancer genome anatomy project: building an annotated gene index. Trends Genet. 2000 Mar;16(3):103-6.
6. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF. GenBank. Nucl. Acids Res. 1998; 26; 1-7.
7. Lehväslaiho H, Ashburner M and Etzold T. Unified access to mutation databases. Trends in Genetics, 1998, 14:5:205-206.
8. Foster NL, Gombosi E, Teboe C, Little RJ. Balanced centralized and distributed database design in a clinical research environment. Stat Med. 2000 Jun 15-30;19(11-12):1531-44.
9. McCray AT, Ide NC. Design and implementation of a national clinical trials registry. J Am Med Inform Assoc. 2000 May-Jun;7(3):313-23.
10. Berman HM, The past and future of structure databases. Current Opinion in Biotechnology 1999, 10:76-80.
11. Sager N, Lyman M, Bucknall C, et al. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994 Mar-Apr;1(2):142-60
12. Moore GW, Berman JJ. Automatic SNOMED coding. Proc Annu Symp Comput Appl Med Care 1994;:225-9
13. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Methods Inf Med 1998 Nov;37(4-5):334-44
14. Grizzle W, Grody WW, Noll WW, Sobel ME, Stass SA, Trainer T, Travers H, Weedn V, Woodruff K. Recommended policies for uses of human tissue in research, education, and quality control. Ad Hoc Committee on Stored Tissue, College of American Pathologists. Arch Pathol Lab Med. 1999 Apr;123(4):296-300
15. Wertz DC. Archived Specimens: A Platform for Discussion. Community Genetics 1999, 2:2–3:51-60.
16. Rashbass, J. The impact of information technology on histopathology. *Histopathology. 36(1):1-7, January 2000*
17. Collaborative Computational Project. <http://www.hgmp.mrc.ac.uk/CCP11/protnucdb.txt.html>
18. Murphy SN, Morgan MM, Barnett GO, Chueh HC. Optimizing healthcare research data warehouse design through past COSTAR query analysis. Proc AMIA Symp. 1999;:892-6.

19. Dubey AK, Chueh HC. Using the extensible markup language (XML) in automated clinical practice guidelines. Proc AMIA Symp. 1998;:735-9.
20. Chueh HC, Raila WF, Berkowicz DA, Barnett GO. An XML portable chart format. Proc AMIA Symp. 1998;:730-4.
21. Chueh HC, Barnett GO. Client-server, distributed database strategies in a health-care record system for a homeless population. J Am Med Inform Assoc. 1994 Mar-Apr;1(2):186-98.
22. Berkowicz DA, Barnett GO, Chueh HC. Component architecture for web based EMR applications. Proc AMIA Symp. 1998;:116-20.
23. Confidentiality, Data, Security and Cancer Research: Report of a Workshop; National Institutes of Health, Bethesda, Maryland; February, 2000.
24. Berman JJ, Moore GW, Hutchins GM. US Senate Bill 422: The Gentle Confidentiality and Nondiscrimination Act of 1997. Diagnostic Molecular Pathology 7(4): 192-196, 1998.
25. Tarczy-Hornoch P, Shannon P, Baskin P, Espeseth M, Pagon RA. GeneClinics: a hybrid text/data electronic publishing model using XML applied to clinical genetic testing. J Am Med Inform Assoc. 2000 May-Jun;7(3):267-76.
26. de Groen PC, Barry JA, Schaller WJ. Applying World Wide Web Technology to the Study of Patients with Rare Diseases. *Annals of Internal Medicine. 129(2):107-113, July 15, 1998*
27. Grizzle WE, Woodruff KH, Trainer TD. The Pathologist's Role in the Use of Human Tissues in Research-Legal, Ethical, and Other Issues. Arch Pathol Lab Med 120: 909-912, 1996.
28. Merz JF, Sankar P, Taube SE, LiVolsi V. Use of Human Tissue in Research: Clarifying Clinician and Researcher Roles and Information Flows. J Investigative Medicine 45(5): 252-257, 1997.

29. Fraser, H. S., I. S. Kohane, et al. (1997). "Using the technology of the world wide web to manage clinical information." British Medical Journal 314(7094): 1600-1603.
30. Hinds, A., P. Greenspun, et al. (1995). WHAM! A forms constructor for medical record access via the world wide web. Proceedings, Annual Fall Symposium of the American Medical Informatics Association, New Orleans, LA, Hanley & Belfus, Inc.
31. Kohane, I., P. Greenspun, et al. (1995). W3-EMRS: Access to Multi-Institutional Electronic Medical Records via with World Wide Web. Spring Congress of the American Medical Informatics Association., Boston, MA.
32. Kohane, I. S. (1996). "Exploring the functions of World Wide Web-based electronic medical record systems." MD Computing 13(4): 339-346.
33. Kohane, I. S., H. Dong, et al. (1998). Health Information Identification and De-Identification Toolkit. Proceedings, Annual Fall Symposium of the American Medical Informatics Association, Florida, Hanley and Belfus, Inc.
34. Kohane, I. S., P. Greenspun, et al. (1996). "Building National Electronic Medical Record Systems via the World Wide Web." Journal of the American Medical Informatics Association 3(3): 191-207.
35. Kohane, I. S., P. Greenspun, et al. (1995). "Accessing Pediatric Electronic Medical Record Systems via the World Wide Web." Pediatric Research 37: 139A.
36. Kohane, I. S., F. J. v. Wingerde, et al. (1996). Sharing electronic medical records across multiple heterogeneous and competing institutions. Proceedings, Annual Fall Symposium of the American Medical Informatics Association, Washington, DC, Hanley & Belfus, Inc.
37. Rind, D. M., I. S. Kohane, et al. (1997). "Maintaining the Confidentiality of Medical Records Shared over the Internet and World Wide Web." Annals in Internal Medicine 127(2): 138-141.
38. Riva, A., K. Mandl, et al. (2000). "The Personal Internetworked Notary and Guardian." International Journal of Medical Informatics In press.
39. Sun, Y., F. J. v. Wingerde, et al. (1999). "The challenges of automating a real-time clinical practice guideline." Clinical Performance and Quality Health Care 7(1): 28-35.
40. Szolovits, P. and I. Kohane (1994). "Against simple universal health identifiers." Journal of the American Medical Informatics Association 1(4): 316-319.

41. Wang, K., I. S. Kohane, et al. (1996). A Real-Time Patient Monitoring System on the World-Wide Web. Proceedings, Annual Fall Symposium of the American Medical Informatics Association, Washington, DC, Hanley & Belfus, Inc.
42. Wang, K., F. J. van Wingerde, et al. (1997). "A Java-based multi-institutional medical information retrieval system." Proceedings, Annual Fall Symposium of the American Medical Informatics Association: 538-42.
43. Wingerde, F. J. v., J. Schindler, et al. (1996). Using HL7 and the World Wide Web for unifying patient data from remote databases. Proceedings, Annual Fall Symposium of the American Medical Informatics Association, Washington, DC, Hanley & Belfus, Inc.
44. Wingerde, F. J. v., Y. Sun, et al. (1998). Linking Multiple Heterogeneous Data Sources to Practice Guidelines. Proceedings, Annual Fall Symposium of the American Medical Informatics Association, Florida, Hanley and Belfus, Inc.
45. Halamka, J. D. and C. Safran (1997). "Virtual consolidation of Boston's Beth Israel and New England Deaconess Hospitals via the World Wide Web." Proceedings, Annual Fall Symposium of the American Medical Informatics Association: 349-53.
46. Halamka, J. D., P. Szolovits, et al. (1997). "A WWW implementation of national recommendations for protecting electronic health information." J Am Med Inform Assoc **4**(6): 458-64.
47. Dolin RH, Rishel W, Biron PV, Spinosa J, Mattison JE. SGML and XML as interchange formats for HL7 messages. Proc AMIA Symp. 1998;:720-4.
48. Sokolowski R, Dudeck J. XML and its impact on content and structure in electronic health care documents. Proc AMIA Symp. 1999;:147-51.
49. Komorowski HJ, Greenes RA. The use of fish-eye views for displaying semantic relationships in a medical taxonomy. Proc Eleventh Annual Symposium on Computer Applications in Medical Care (SCAMC); Washington, DC. New York: IEEE Computer Society Press. November, 1987; 113-116
50. Appel RD, Komorowski HJ, Barr CE., and Greenes RA. Intelligent focusing in knowledge indexing and retrieval the relatedness tool. Proc Twelfth Annual Symposium on Computer Applications in Medical Care (SCAMC), Washington, DC. New York: IEEE Computer Society Press. November, 1988; 152-157
51. Barr CE, Komorowski HJ, Pattison-Gordon E, Greenes RA. Conceptual modeling for the Unified Medical Language System. Proc Twelfth Annual Symposium on Computer Applications in Medical Care (SCAMC), Washington, DC. New York: IEEE Computer Society Press. November, 1988; 148-151
52. Hersh WR, Greenes RA. Information retrieval in medicine: State of the art. MD Comput 1990; 7(5): 302-311
53. Hersh WR, Greenes RA. SAPHIRE An information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Comput Biomed Res 1990; 23(5): 410-425
54. Hersh WR, Pattison-Gordon E, Greenes RA. Adaptation of Meta-1 for SAPHIRE, a general purpose information retrieval system. Proc Fourteenth Annual Symposium on Computer Applications in Medical Care (SCAMC), Washington, DC. New York: IEEE Computer Society Press. November, 1990; 156-160
55. Bell DS, Greenes RA, Doubilet PD. Form-based clinical input from a structured vocabulary: Initial application in ultrasound reporting. Proc Sixteenth Annual Symposium on Computer Applications in Medical Care (SCAMC), Baltimore, MD, Nov 92. New York: McGraw-Hill. 1993; 789-790
56. Bell DS, Pattison-Gordon E, Greenes RA. Experiments in concept modeling for radiographic image reports. JAMIA, 1994; 1(3): 249-262
57. Pattison-Gordon E, Greenes RA. An empirical investigation into the conceptual structure of chest radiograph findings. Proc Eighteenth Annual Symposium on Computer Applications in Medical Care (SCAMC), Washington, DC. Nov, 94. Philadelphia: Hanley & Belfus. 1994; 257-261

H. Consortium/Contractual Arrangements

See attached. Contractual letters are from UCLA Healthcare, Cedars-Sinai Medical Center and DFCI.
See attached. Agreement, in principal, to share specimens and information:

1. Children's Hospital, Boston, MA

2. Massachusetts General Hospital, Boston, MA
3. Beth Israel Deaconess Medical Center, Boston, MA
4. Brigham & Women's Hospital, Boston, MA
5. Cedars-Sinai Medical Center, LA, CA
6. Olive View Medical Center

I. Consultants

None

Appendix 1: Letters of support

Appendix 2:
Narrative description of information systems at consortium institutions

Massachusetts General Hospital (MGH) Pathology Information System:
MGH pathology laboratory information system was a homegrown MUMPS system originally written in 1976. The department converted its laboratory information system to CoPath in 1990, which interfaces with the MGH Patient Care Information System (PCIS). As part of the consolidated Information Systems Department of Partners HealthCare System, PCIS data is linked with larger information repositories, such as the new Longitudinal Medical Record database which consolidate patient clinical information from across all of the Partners affiliated entities. (Partners HealthCare System includes BWH, MGH and a number of local hospitals) Given the extremely large size of the patient database in the Partners system, it is an excellent resource for clinical research project development. There is a total of approximately 1.2 millions cases online in the MGH CoPath database. In the decade of the 1990's, there are an estimated total of 940,000 pathology cases, 55% of them are surgical pathology cases with archived specimens.

Brigham and Women's Hospital (BWH) Pathology Information System:
The Brigham Information and Computing System, BICS, is a homegrown system using MUMPS-based programming language. BICS is an integrated system at BWH incorporating a wide variety of functions, including: on-line patient medical record, which incorporates all patient reports (e.g., admitting notes, discharge abstracts, operative notes, laboratory results, etc.), patient registration data, on-line test-ordering, pharmacy system, etc. As a member of the Partners HealthCare System, BICS data is also linked with larger information repositories across all of the Partners affiliated entities. The BWH Pathology laboratory information system is comprised of a number of individual modules of the integrated BICS system at BWH. There are now 4 discrete modules: surgical pathology, cytology, autopsy, and cytogenetics. A. Surgical pathology: The system was operational in 1987, and between 1987 & 1989, there were 89,225 cases entered into the system. Currently, only the final diagnosis information is available on-line, while the gross description and other clinical information on each case has been purged. From 1990 to 1999 there were 426,637 cases with full reports available on-line. B. Cytology: The system was operational in 1988, and between 1988 & 1989, there were 82,599 cases on-line, with full reports still available. From 1990 to 1999 there were 530,991 cases with full reports on-line. C. Autopsy: The system was operational in 1989, with 402 reports available on-line. From 1990 to 1999 there were 3,360 autopsy cases, with full report on-line. D. Cytogenetics: The system was operational in 1998, and between 1998 & 1999 there were 14,581 cases, with full reports on-line. For the decade of the 1990s, there were a total of 975,569 pathology cases with full reports on-line in BICS. 83% surgical pathology cases have archived specimens, i.e. 47% of total cases.

Beth Israel Deaconess Medical Center (BID) Pathology Information System:
Beth Israel Hospital merged with Deaconess Medical Center in 1996 to form BID. BID East Campus (formerly Beth Israel Hospital) Center for Clinical Computing, CCC hospital information system was originally written in 1970s-80s. The original language was MIIS and it has since been translated into MUMPS. The surgical pathology information module has been in use since 1984. BID west campus (formerly Deaconess Hospital) used a commercial system, CHC, from 1990 to 1999. Data from the CHC system exists currently as stand alone text files in an Access database, which could be used to search for individual reports. As of July 1, 1999 all cases within BID are accessioned and reported in CCC. The CCC system was purged in 1997. From 1995-1996, there are about 50,000 cases with partial case data still available online. These reports have final diagnosis and addenda, but are missing at least the gross description and clinical information fields. Full report data still online from 1997. There are 85,000 pathology cases between 1997-1999. 93% are surgical pathology cases and they all have archived specimens. Autopsy reports exist only on paper as word files with associated archived specimens. Cytology reports should be similar to pathology. Approximately 45,000 cases per year currently.

Children's Hospital (CH) Pathology Information System

Children's Hospital first utilized an electronic pathology reporting system in 1992. It is a commercial laboratory system: Path Net Laboratory system, AP module, Cerner Corporation. The system utilizes RMS files, and is interfaced with the various Hospital information systems. The Hospital uses an Oracle Database to handle all clinical information, including pathology results. The information is protected by a high level of security, with a fire-wall preventing access from the outside. Retrieving information requires multiple levels of security, including series of passwords, and random-number generators (SecureID) for off-site retrieval of information. A user-friendly application is currently in piloting use (Results Reporting), which posts all information from the Oracle

Appendix 2: continued

Database clinical repository in a web-based format (intranet), including pathology results, other laboratory results, radiology information, and clinic visits/operating reports. Laboratory results are retained within the Cerner System for varying amounts of time. Pathology results (especially the clinically significant fields) are retained within the Cerner System indefinitely. Some information is purged from the system at defined periods, such as laboratory workloads, QC data, date and time markings of laboratory actions. Pathology results within the Oracle Database include only the significant data related to the report, and is retained in this database indefinitely. Estimating the total volume of cases with archival tissue as of July 1, 2000, in the laboratory database: 55,000

University of California at Los Angeles (UCLA) Pathology Information System

The laboratory information service (LIS) system is Meditech and has been online since 1994. From 1988 through 1994 the LIS was a CHC system. The Pathology information system is CoPath and has been online since 1994. From 1989 through 1994 the pathology was reported on the CHC system. The hospital demographics information system (HIS) is an in-house developed system. A hospital clinical repository including medical histories, radiology reports, laboratory test results, and anatomic pathology also resides on the hospital mainframe. The clinical data is accessible by mainframe terminals or on an in-house developed WEB based system. The interface from Meditech LIS to the data repository is TCP/IP, HL7 and passes through a Datagate report engine. The interface from CoPath to the data repository is TCP/IP, custom format and also passes through a Datagate report engine. Full pathology and laboratory report data is available via the ancillary system or the mainframe clinical repository from 1989 to the present.-CoPath system since late 1994; CHC before that with full reports of cases from 1989 on transferred to CoPath
Total cases online: 470,263; Archived tissue for online cases: 233,736; Estimate of cancer cases: 20,000
1999 statistics: 64,832 total 33,800 surgical. Full report data available for surgical and cytology from 1989; autopsy from 1998.

Cedars-Sinai Medical Center (CSMC) Pathology Information System

The laboratory information service (LIS) system is Sunquest and has been online since 1987. The hospital demographics information system (HIS) is an in-house modified ADS-Plus.The HIS is scheduled to be replaced by our own site-developed software by mid 2002. A hospital clinical repository called WEB VS including medical histories with facesheets, radiology reports, laboratory test parameters, and anatomic pathology reports is on a VAX VMS system that is web accessible. This data repository is currently being migrated to an Oracle database. The interface between Sunquest and the data repository is TCP/IP, HL7 and passes through a Datagate report engine. Full pathology report data is available via Sunquest from 1987 to the present and on the WEB VS from 1994 to the present. Total cases online (1990-99): 541,008
Archived tissue for online cases: 273,508. Estimate of cancer cases: 69,225
1999 statistics:75,575 total 35,400 surgical; 40,000 cytology; 175 autopsy.

VA Greater Los Angeles Healthcare System (VAGLAHS) Pathology Information System

The Vista or DHCP is a decentralized, relational database that is supported both nationally and locally, This system utilizes MUMPS programming language. Vista is an integrated hospital Information system at VAGLAHS incorporating a wide variety of functions, includingi on-line patient medical record. which

incorporates all patient reports (e.g,, admitting notes, discharge abstracts, operative notes, laboratory results), patient identification and demographic information, on-line test-ordering, and pharmacy system. The database at VAGLAHS includes consolidated 'information from three large ambulatory care centers, ten community based outreach clinics, and one tertiary care facility, representing about 75,000 patients. There are date links to four southwestern U.S. healthcare facilities, VAGLAHS is developing a number of web-based applications to further enhance the functionality of Vista. The VAGLAHS Pathology information system is comprised of a number of closely-related modules of the integrated Vista system. There are now 4 discrete modules: surgical pathology, cytology, and autopsy. VISTA system operating since 1991. Total cases online: (1991-2000): 93,907 Archived tissue for cases online: 58,201

Olive View Medical Center (OVMC) Pathology Information System

The Olive View-UCLA Medical Center Pathology Information System is a component of the HIS (Compucare Affinity) since 1999. Prior records are kept on a PC based database. Modifications to the Compucare HIS were developed in September 1999 that included the installation of GUI32 overlay. This modification provided full text results entry using MS Word of AP into the HIS through a sub-module (Department Management). Now

Appendix 2: continued

complete AP results are available for inquiry and electronic reporting through the HIS. This is a separate system from the LIS and the patients AP results are not linked to Clinical Laboratory data. The Laboratory Information System is Compucare / Sigma system which uses MEESE Data Base to support all Clinical information (General Lab., Serology, Microbiology, Special Chemistry, Flow Cytometry, and Reference / Send-Out Labs.). Component of Compucare Affinity HIS since 1999; prior pathology data maintained on a personal computer database.
Total surgical cases since 1990: 81,040
Estimate of cancer cases: 7,000
1999 surgical cases: approx. 7000
Full report data available since 1990
Archived slides and blocks: 25+ years

Appendix 3:
Composite data element set for pathology information systems in the consortium:

Data Elements stored/accessible in MGH CoPath—all fields are searchable:

Accession number
Patient Demographics:
- Name
- Unit number (Medical Record Number = MRN)
- Age as Date of Birth (DOB)
- Sex

Specimen submitted—comes in part, each of which corresponds to ONE specimen
Data of procedure
Date specimen received
Date of report—signed out
Grossing doctor—Resident
Diagnosing doctor
Prior specimen numbers (no limit)
Diagnosis (text field)
Gross description (text field)
Clinical history (text field)
Charges—in CPT-4 codes
SNOMED code (ok w/accuracy) and ICD-9, not linked
ICD-9 codes include:
- Clinical Dx code
- No Clinical history
- Intra-operative (Frozen Sections) Dx

Addenda (for special studies)

Appendix 4.

Additional Information about DFCI CRIS and STIP

Clinical Research Information System (CRIS)

The clinical data collection strategies for CRIS were first developed for breast cancer, and all other diseases have followed that paradigm. For patients with breast cancer, data are entered directly into CRIS for any patient receiving some or all of her breast cancer care at the participating institutions. Patients seen for one-time-only consultation contribute patient self-reported data at presentation (including family history and other risk factor data), but baseline clinical data and follow-up data are not collected. The decision about whether to classify a patient as a consult is made 3 months after the first visit to ensure that patients who are listed as consults initially, but who then decide to receive their care at the institution, are included. Data are collected for CRIS at the time of the patient's first presentation to the institution and in follow up. Patients who transfer their care out of the DF/HCC are followed for vital status only. In brief, the data elements collected include:

Intake:

Patient Self-Reported Data: sociodemographics, triggering event, family history, smoking and alcohol history, ob-gyn history, past cancers, comorbidity, performance status, days in hospital attributable to breast cancer, and days lost from work attributable to breast cancer.
Staging: diagnosis date, TNM staging, metastatic sites, prior treatment, response to prior treatment, current disease status, and current treatment status.
Pathology: pathology/histology of all specimens.

Follow Up:

Patient Self-Reported Data: menopausal status, employment status, performance status, days in hospital attributable to breast cancer, days lost from work attributable to breast cancer, and patient satisfaction.

Medical Follow Up: treatments, response to treatment, symptoms, complications, recurrence, treatment status, disease status, and vital status.

The general strategy for data collection is to obtain information directly from patients using waiting-room surveys for those data elements that can only reliably be obtained from patients. This information is immediately entered into CRIS by a dedicated research assistant (RA) in the clinic, and the results tabulated and summarized on a printout that is attached to the chart for use by providers during that clinic visit. Physicians are provided with pre-printed, patient-specific forms on which to record medical data (exclusive of pathology) at the time of each patient visit. Medical data on patients is entered into CRIS by data managers who use the physician forms, the paper chart, the electronic medical record, and the tumor registry as data sources.The eventual collection of complete medical data is not dependent on completion of forms by physicians, since the records of all patients are fully abstracted by data managers. Therefore, there is no requirement that physicians complete either new patient or follow-up forms.
Use of CRIS data to support research and administrative decision-making is facilitated by a user-friendly query tool, programmed in Business Objects. This application allows the user to "ask questions" of the database in a intuitive fashion, as well as to run previously programmed, customized reports.

Specimen Tracking Information Program (STIP)

STIP is a discrete but compatible module of the Oracle database designed to track specimens for banking and research use. Many of the features described above for CRIS also apply to STIP, including the user-friendly Power Builder data entry screens, relational database structure, and ability to run reports through Business Objects. STIP references the CRIS patient table; it does not maintain its own patient database. But it does include a registration function to allow the capture of data for patients who contribute blinded specimens and/or those who are not receiving care in the institution and are therefore not included in CRIS. In addition, users

may access CRIS data on a read-only basis for patients whose specimens are logged into STIP. This access is controlled through unique individual-level password protection.

STIP includes data entry functions to capture data for specimens, components of specimens, orders to collect or send out specimens, patient-level data associated with specimen collection, research tests, and research studies. In addition, there are administrative data entry functions to maintain tables of contacts, storage repositories, families, code tables used in the application, and research studies. Specimen level data include: patient; specimen-contents; dates collected and received; provider's name; nature of the individual stored components (e.g. slides or aliquots); specimen-type; amount; date-stored; storage-location; storage-status (e.g. stored/in-preparation/sent out/destroyed); send out date; and sender's name. Data elements relevant to orders include: type of order (collection or send out); lists of patients; lists of tests (for send out orders); type of action ordered; protocol, authorizing MD; individual sending/receiving the specimen; and workflow dates (entry-date, action-date, action-taken-date, order-filled-date). The system tracks orders with pending-order work lists. Detailed tracking and results data may be entered for any tests performed on the specimen. Repository level data (contents of specific storage repositories such as freezers and slide cabinets, and divisions of those repositories) are also available in the system.

Security is maintained with two layers of individual-level password access. All users must have both a network password and a STIP password.Passwords are linked to specific job categories -- individual users have access only to those functions relevant to their jobs. The system also allows for multiple security groups, each with its own specimen inventory.

Both **CRIS** and **STIP** employ a multi-user, client-server architecture with Windows client machines and a Unix Oracle database server.

Appendix 5

Information on Informed Consent Practices for Research:

For studies that entail minimal risk to participants, the informed consent document should seek as broad an authorization from participants as possible, provided that broad consent is consistent with imparting the necessary information effectively to participants. This is also the most convenient approach, since it tends to obviate the need to go back to study participants for more specific or follow-up consents at a later date. For example, it may be very problematic logistically to obtain consent for additional follow-up information for long-term epidemiological studies or clinical trials or for additional information to annotate archived samples. General consent for future minimal-risk epidemiologic research can often be obtained by providing an option in the consent stating explicitly that "you do not need to contact me." This allows tissue or other health information to be used in future minimal-risk research without re-contacting the subject. The alternative option that should also be provided is "I would like to be contacted in the future."

On the other hand, for many complex studies, the informed consent is better structured and presented to participants in a modular fashion. Particularly in studies wherein participants are offered multiple requests or procedures, a staged consent process, with the opportunity to consent to each request separately, may more effectively inform and educate participants than a single, complex, multi-purpose form which may be confusing. Another advantage is that the study is less likely to be compromised; a potential participant might, for example, be agreeable to a questionnaire but not the drawing of blood. The potential for re-contact to secure permission for future uses of additional information or specimens should be specifically stated. A variation on this theme is the use of a tiered consent form that offers participants the choice of whether their data or tissue will be used solely for the original research intent or for purposes in the future that are currently unanticipated. Any of these options enables greater control by study participants over the future use of their biologic sample and research information. Such an approach honors the autonomy of participants by making their preferences specific.

It is highly desirable to avoid the need to obtain consent repeatedly from participants simply because of the passage of time. The need for frequent and repeated consent to re-authorize unchanged study objectives is demeaning to study participants, does not enhance participant autonomy, and is a significant barrier to the conduct of research. Participants should, of course, always have the right to withdraw from further participation in research at any time. Consent forms should clearly state that the participant can withdraw consent at any time, rather than specifying an expiration date for the consent. Expiration dates patronize the research participant by implicitly questioning his/her decision-making capacity.

Whatever their form and structure, consent forms should inform participants that participation in a research study means that the research team will have access to their records. The team may, for example, include physicians or other investigators, research nurses, data managers, and pharmacists. Once confidentiality training is the norm, the consent should note that these individuals have had confidentiality training and are bound by the confidentiality policies of the institution. Patients should also be notified about the extent to which their records will be accessible to officials from the FDA, NCI, drug or device manufacturers tat may be sponsoring the study, or other groups. Study participants should be advised that all efforts will be made to keep their research data confidential, but that their data remain subject to the subpoena power of the courts, in the event of pertinent legal action.

A rather subtle aspect of the relationship between informed consent and participant privacy relates to the possibility of identifying study participants inadvertently from published analyses of study data. This is ordinarily not an issue because of the impossibility of identifying individuals from summary data. If, however, analysis of a particular study involves such small subsets or focuses on a rare enough condition that patients are highly likely to be identifiable, then analysis of the data should require IRB approval, Similarly, publication of

results involving an individual or small subset of patients with unique distinguishing features such as pedigrees, which would make them highly likely to be identifiable, should require IRB approval prior to publication.

For tissue specimens, informed consent authorizing research use should be on record for all specimens and accompanying clinical data. This is not usually a problem when the specimen has been procured specifically for a particular research study or for entry into a research archive, but it can be difficult in the routine setting of clinical pathology, where administrative arrangements for keeping research-related records often do not now exist. It would be most efficient to combine consent for research use of tissue and associated data with the general surgical consent but to include a separate signature line permitting research use. Whenever it is anticipated that tissue will be sampled or removed for diagnosis or treatment, consent should, if possible, be obtained at the time of admission for in-patient and out-patient procedures, and this consent should be recorded and tracked. The actual process for obtaining consent should be left up to the physician or the institution. Certain patients may be too distracted or upset to carefully review and consider the content of these documents at the time they are presented. Some practitioners give a copy of the consent document, or a transcript of the procedures-and-risks discussion, to take home to review at leisure prior to the actual signing of the consent.

When archiving new tissues with their clinical annotations, the informed consent for using them in unspecified research involving minimal risk should be prospective and as broad as possible, so that it will not be necessary to go back to the patient for more specific consent at some later date. The same is true for the use of medical records or other information repositories for future minimum risk research. IRBs should be strongly encouraged not to require re-consent to authorize minimal-risk research.

When patients refuse or withdraw consent for the research use of their specimen, these should be annotated accordingly and not used for research. It should be possible for patients to give or refuse consent for the use of clinical data separately from the use of tissue. The default position should be “no”—that is, the sample and data cannot be used for research unless there is a positive record of consent.

Archived specimens collected prior to the current standards for informed consent should be available for research proposed, as long as appropriate steps are taken to maintain confidentiality of identifiable information. The decision regarding whether the specimens may be used should be in accord with the Common Rule definition of risk, allowing waiver of the need for consent in research that is determined by an IRB.

General guidelines for all requests for specimens:

- A review and evaluation of the protocol is necessary to protect against the use of identifiable research data when non-identifiable data would suffice; to establish that the requestor is qualified to perform the proposed research and will employ reliable methodologies; to establish that the requestor is capable of protecting confidentiality; and to prevent uses of the data for purposes other than those for which they were collected without any necessary additional review.
- The investigator and the party releasing the identifiable data should sign a binding research agreement. The agreement should specify the terms under which the data may be used and how confidentiality must be maintained, including sanctions or penalties for breach of confidentiality. It should prohibit re-release of identifiable data to third parties, and should delineate any other obligations of the requesting investigator. In addition to the agreement, the party releasing the data may provide written documentation of recommended policies and practices for using the data, including relevant legal issues. Such material may help prevent inadvertent breaches.
- There should be evidence of external review and approval of the data release. An approved IRB or the equivalent should perform the review.
- The researcher receiving the data or specimens should be able to provide documentation that all members of the receiving research group have been trained in confidentiality practices.